On swarm algorithms

Laurent Miclo

Toulouse School of Economics
Institut de Mathématiques de Toulouse

Joint work with Jérôme Bolte and Stéphane Villeneuve

Plan of the talk

Plan

Global optimization

On a compact Riemannian manifold M, consider a smooth function $U : M \to \mathbb{R}$. We would like to find its global minima, or at least a point close to their set

$$\mathcal{M} \quad := \quad \{x \in M \, : \, U(x) = \min_{M} U\}$$

Simulated annealing is one of the few general algorithms carrying out this task. It is a time-inhomogeneous diffusion which tends to concentrate around $\mathcal{M}$ in long time.
In practice its slow convergence can be enhanced by sending several independent particles following the same evolution. Our goal here is to generalize this procedure by considering interacting particles.

Simulated annealing

Heuristically simulated annealing can be described particle on M evolving according to the sde

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

where

• $(B(t))_{t\geqslant 0}$ is a Brownian motion on M (with generator the Laplacian $\triangle/2$),

• $(\beta_t)_{t\geqslant 0}$ is an evolution of the inverse temperature, which is taking non-negative values, is non-decreasing and diverges to $+\infty$ in large time.

Thus $(X(t))_{t\geqslant 0}$ is a random perturbation of the gradient descent associated to U, whose relative strength is increasing with time.

Convergence of simulated annealing

Consider logarithmic evolutions:

$$\forall\ t \geqslant 0, \qquad \beta_t \quad = \quad k^{-1} \ln(1+t)$$

It can be shown there exists a critical constant $c \geqslant 0$ such that if $k \geqslant c$, for any neighborhood $\mathcal{N}$ of $\mathcal{M}$,

$$\forall\ \epsilon > 0, \qquad \lim_{t \to +\infty} \mathbb{P}[X(t) \in \mathcal{N}] \quad = \quad 1$$

but this convergence is wrong if $k < c$. We have $c > 0$ as soon as there is a least one local minimum which is not global.

Time-inhomogeneous ergodicity

This result has been obtained by several approaches: direct renewal computations, large deviations, functional inequalities. The latter method first proves time-inhomogeneous ergodicity: there exists another critical constant $c' \geqslant 0$ such that if $k \geqslant c'$, in total variation,

$$\lim_{t \to +\infty} \|\mathcal{L}(X(t)) - \pi_{\beta_t}\|_{\mathrm{tv}} \quad = \quad 0$$

where $\mathcal{L}(X(t))$ is the law of $X(t)$ and π_{β_t} is the instantaneous invariant measure at time $t \geqslant 0$. It is given by the Gibbs measure associated to the potential U at temperature $1/\beta_t$, namely the density of π_{β_t} with respect to the Riemann measure ℓ is proportional to

$$\exp(-\beta_t U)$$

In general $c' \geqslant c$ and there is equality for instance if there is a unique global minimum.

Speeding-up?

Unfortunately, these convergences are quite slow, due to the logarithmic feature of the inverse temperature.
In practice, faster temperature schemes are considered with finite time horizons. Alternatively, independent simulations $(X_1(t))_{t\geqslant 0}$, $(X_2(t))_{t\geqslant 0}$, ..., $(X_N(t))_{t\geqslant 0}$ of $(X(t))_{t\geqslant 0}$ are used.

Is it possible to introduce interactions between these particles to improve the speed of convergence?

The goal of this talk is to present such an interacting system, where each particle will boost its Brownian motion when there are too few or too much particles surrounding it.

Plan

The pde point of view

For any $t \geqslant 0$, denote $\rho(t)$ the law of $X(t)$.
By ellipticity, it can be seen that for any $t > 0$, $\rho(t) \ll \ell$ and the probability measure ρ is identified with its density with respect to ℓ.
The probability-measure dynamical system $(\rho(t))_{t \geqslant 0}$ is solution to the parabolic equation

$$\forall\ t > 0, \qquad \dot{\rho}(t) \quad = \quad \beta_t \operatorname{div}(\rho(t) \nabla U) + \triangle \rho(t)$$

where div, ∇ and $\triangle$ are the divergence, the gradient and the Laplacian associated to the Riemannian structure of M.

In Wasserstein space

Via Otto's formalism, the flow $(\rho(t))_{t\geqslant 0}$ is a gradient descent in the Wasserstein space $\mathcal{W}$:

$$\forall\ t \geqslant 0, \qquad \dot{\rho}(t) \quad = \quad -\mathrm{grad}_{\mathcal{W}}\, \mathcal{U}_{\beta_t}[\rho(t)]$$

where $\mathcal{U}_\beta$ is relative entropy of $\rho(t)$ with respect to the Gibbs measure π_β, and where the gradient is with respect to the infinite-dimensional Riemannian-like structure of $\mathcal{W}$.
Recall that $\mathcal{W}$ is the space of probability measures on M equipped with the Monge-Kantorovich distance defined through

$$W_2^2(\mu,\nu) \quad = \quad \inf\left\{ \int_M \delta^2(x,y)\, p(dx,dy) \ :\ p \in \mathcal{C}(\mu,\nu) \right\}$$

where δ is the Riemannian distance on M and $\mathcal{C}(\mu,\nu)$ is the set of couplings on M^2 of the two probability measures μ and ν on M.

Relative entropy

Up to an additive constant, the relative entropy can be expanded as

$$\mathcal{U}_\beta[\rho] \quad = \quad \beta \int_M U \, d\rho + \int_M \rho \log \rho \, d\ell$$

at least for $\rho \ll \ell$.

The term $\int_M U \, d\rho$ can be seen as an up-lift to $\mathcal{W}$ of the function U and $\int_M \rho \log \rho \, d\ell$ is a penalization term. The inverse temperature β enables to tune their relative importance.
Due to the gradient descent structure, $\mathcal{U}_\beta$ can serve as a Liapounov functional for the flow $(\rho(t))_{t \geqslant 0}$, at least when the inverse temperature is fixed. The comparison between $\mathcal{U}_\beta[\rho(t)]$ and the "entropy production" term $\frac{d}{dt}\mathcal{U}_\beta[\rho(t)]$ uses logarithmic Sobolev inequalities. Its adaptation to the time-inhomogeneous case leads to the ergodicity result, because $\mathcal{U}_{\beta_t}[\rho(t)]$ enables us to control

$$\|\rho(t) - \pi_{\beta_t}\|_{\mathrm{tv}}$$

Penalizations (i)

Consider another penalization term:

$$\mathcal{H}[\rho] \quad = \quad \begin{cases} \displaystyle\int_M \varphi(\rho)\, d\ell & \text{if } \rho \ll \ell \\ +\infty & \text{otherwise,} \end{cases}$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly convex function satisfying $\varphi(1) = 0$ and smooth on $(0, +\infty)$. For the relative entropy, it corresponds to

$$\forall\ r \geqslant 0, \qquad \varphi_1(r) \quad := \quad r \ln(r) - r + 1$$

Another traditional function, for $m > 0$, $m \neq 1$:

$$\forall\ r \geqslant 0, \qquad \varphi_m(r) \quad := \quad \frac{r^m - 1 - m(r-1)}{m(m-1)}$$

Penalizations (ii)

We generalize the relative entropy to

$$\mathcal{F}_\beta[\rho] \quad := \quad \beta \int_M U \, d\rho + \mathcal{H}[\rho]$$

As we are to see, for any $\beta > 0$, this functional has a unique local minimizer μ_β, contrary to U, as soon as $\varphi'(0) = -\infty$. Furthermore this minimizer concentrates around $\mathcal{M}$ for large β.

It is then natural to consider the corresponding (time-inhomogeneous) gradient descent in $\mathcal{W}$ and to investigate its probabilistic translations into interacting particle algorithms.

Plan

Back to pde

For a while, let us fix the inverse temperature $\beta \geqslant 0$. For the generalized relative entropy, we compute that

$$\operatorname{grad}_{\mathcal{W}} \mathcal{F}_\beta[\rho] \quad = \quad -\operatorname{div}(\rho(\beta \nabla U + \nabla \varphi'(\rho)))$$

and we are led to the non-linear evolution equation

$$\dot{\rho} \quad = \quad \operatorname{div}(\rho(\beta \nabla U + \nabla \varphi'(\rho))) \tag{1}$$

A corresponding stationary measure μ_β satisfies

$$\operatorname{div}(\mu_\beta(\beta \nabla U + \nabla \varphi'(\mu_\beta))) \quad = \quad 0 \tag{2}$$

Integrating by parts, it appears that $\beta U + \varphi'(\mu_\beta)$ has to be constant on each connected component of the support of μ_β.

The minimizer

Lemma 1

Assume $\varphi'(0) = -\infty$, then there exists a unique stationary density μ_β solution to (2). Moreover,

(i) *μ_β is positive everywhere on M and is characterized by the relation*

$$\mu_\beta \quad = \quad \psi(c_\beta - \beta U)$$

where ψ the inverse of φ' and c_β is a normalization parameter characterized by the condition

$$\int_M \psi(c_\beta - \beta U)\, d\ell \quad = \quad 1$$

(ii) *μ_β is the global minimizer of $\mathcal{F}_\beta$.*

(iii) *For any neighborhood $\mathcal{N}$ of $\mathcal{M}$, we have*

$$\lim_{\beta \to +\infty} \mu_\beta[\mathcal{N}] \quad = \quad 1$$

Probabilistic formulation

The integration by parts equally leads to the weak formulation of (1): for any regular test function f,

$$\forall\ t > 0, \qquad \frac{d}{dt}\rho(t)[f] \quad = \quad \rho(t)[L_{\rho(t)}[f]]$$

where

$$L_\rho[f] \quad := \quad \alpha(\rho)\triangle f - \langle \beta \nabla U, \nabla f \rangle$$

with

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

The generator L_ρ depends on ρ through its diffusion coefficient. It leads to non-linear Markov processes whose evolution at any given time depends on the time-marginal. Particle systems can be used to approximate them.

Particular functions φ (i)

Consider the case $U = 0$ and the power-like function φ_m, for $m > 0$, $m \neq 1$. The evolution equation (1) writes

$$\dot{\rho} \quad = \quad \triangle \rho^{m-1}$$

This evolution is well studied and is called the porous media equation for $m > 1$ and the fast diffusion equation for $m < 1$. It was rather investigated when M is an Euclidean space, the long-time behavior is then described by a "convergence" toward a time-renormalized self-similar distribution, called the Barrenblatt solution. This is different from our compact situation. Note that $\varphi'_m(0) = -\infty$ amounts to $m \leqslant 1$.

Particular functions φ (ii)

We will consider functions of the type

$$\forall\ r \geqslant 0, \qquad \varphi_{m,2}(r) \ := \ \begin{cases} \varphi_m(r) & \text{if } r \in (0,1], \\ \varphi_2(r) = \frac{(r-1)^2}{2} & \text{if } r \in (1,+\infty). \end{cases}$$

with $m \in (0, 1/2)$. Note that $\varphi_{m,2}$ is $\mathcal{C}^2$, since for any $m > 0$, $\varphi_m(1) = 0$, $\varphi'_m(1) = 0$ and $\varphi''_m(1) = 1$.
These choices imply

$$\lim_{r \to 0_+} \alpha(r) \ = \ +\infty \ = \ \lim_{r \to +\infty} \alpha(r)$$

which corresponds to the qualitative behavior alluded to above. Furthermore for these functions, there is a unique stationary measure μ_β and we have the estimates:

$$(1 + (1-m)\beta \mathrm{osc}(U))^{\frac{1}{m-1}} \leqslant \min_M \mu_\beta \leqslant \max_M \mu_\beta \leqslant \beta \mathrm{osc}(U) + 1.$$

where $\mathrm{osc}(U) := \max_M U - \min_M U$.

Plan

Generalized entropy production

Denote

$$\begin{aligned}\mathcal{I}[\rho] &:= \mathcal{F}_\beta(\rho) - \min_{\mathcal{W}} \mathcal{F}_\beta = \mathcal{F}_\beta(\rho) - \mathcal{F}_\beta(\mu_\beta) \\ &= \int_M \varphi(\rho) - \varphi(\mu_\beta) - \varphi'(\mu_\beta)(\rho - \mu_\beta)\, d\ell\end{aligned}$$

and consider the evolution equation (1). Since it is a gradient descent, it is natural to investigate the evolution of $\mathcal{I}(\rho(t))$. We compute

$$\frac{d}{dt}\mathcal{I}[\rho(t)] = -\mathcal{J}[\rho(t)]$$

with

$$\mathcal{J}[\rho] = \int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho\, d\ell$$

Comparison of $\mathcal{I}$ and $\mathcal{J}$?

To get a differential inequality satisfied by $\mathcal{I}[\rho(t)]$, we must compare it to $\mathcal{J}[\rho(t)]$. Indeed via Łojasiewicz-type arguments:

Theorem 2

Assume there exist $c(\beta) > 0$, $\Omega : \mathbb{R}_+ \to \mathbb{R}_+$ increasing, such that an inequality of the type

$$\int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu_\beta)|^2 \rho \, d\ell \tag{3}$$
$$\geqslant \quad c(\beta)\,\Omega\left(\int_M \varphi(\rho) - \varphi(\mu_\beta) - \varphi'(\mu_\beta)(\rho - \mu_\beta)\, d\ell\right)$$

holds true whenever ρ is measurable and the left hand side is finite. Then for large $t \geqslant 0$,
(i) $\mathcal{F}_\beta[\rho(t)] \to \mathcal{F}_\beta(\mu_\beta)$.
(ii) If moreover $\Omega(s) = \Theta(s^{2\theta})$ at 0, with $\theta \in (0,1)$, then $\rho(t)$ tends to μ_β for the Monge-Kantorovich metric, i.e. for the weak convergence.

A functional inequality

While we conjecture an inequality such as (3) holds for any compact manifold and for more general functions φ, for the moment it is only proved in a restricted setting:

Theorem 3

Assume M is the circle $\mathbb{T}$ and $\varphi = \varphi_{m,2}$, with $m \in (0, 1/2)$. Then (3) holds with

$$c(\beta) = O\left(\beta^{\frac{-3(2-m)}{1-2m}}\right)$$

$$\Omega(r) = \begin{cases} r^{\frac{3}{2}} & \text{if } r \in [0,1) \\ r^{\frac{1-2m}{2(1-m)}} & \text{if } r \geqslant 1 \end{cases}$$

A corresponding Talagrand-type inequality holds between $\mathcal{I}[\rho]$ and $\mathcal{W}_2(\rho, \mu_\beta)$.

Plan

Time-inhomogeneous mean-field evolution

We come back to a time-dependent inverse temperature scheme $t \mapsto \beta_t$. We restrict our attention to the restricted setting of the previous theorem and we consider the evolution

$$\forall\ t \geqslant 0, \qquad \dot{\rho}(t) \quad = \quad \mathrm{div}(\rho(t)(\beta_t \nabla U + \nabla \varphi'(\rho(t)))) \qquad (4)$$

starting from a given initial distribution $\rho(0)$.
Define $\mathcal{I}[t, \rho]$ and $\mathcal{J}[t, \rho]$ as before, except these quantities now explicitly depend on time through β_t.
We compute that for any time $t > 0$,

$$\frac{d}{dt}\mathcal{I}[t, \rho(t)] \quad = \quad -\mathcal{J}[t, \rho(t)] + \dot{\beta}_t \int_M U(\rho - \mu_{\beta_t})\, d\ell$$

A differential inequality

Writing $v(t) \coloneqq \mathcal{I}[t, \rho(t)]$ and using the functional inequality, we end up with

$$\dot{v}(t) \quad \leqslant \quad -c(\beta_t)\Omega(v(t)) + \operatorname{osc}(U)\left|\dot{\beta}_t\right|$$

Elementary arguments enable to see that

$$\lim_{t \to +\infty} v(t) \quad = \quad 0$$

as soon as

$$\lim_{t \to +\infty} \dot{\beta}(t)/c(\beta(t)) \quad = \quad 0$$

$$\int_1^{+\infty} c(\beta(t))\, dt \quad = \quad +\infty$$

Admissible temperature schemes

The previous conditions are satisfied if we take

$$\forall\ t \geqslant 0, \qquad \beta(t) \ := \ kt^{\gamma}$$

with $k > 0$ and $\gamma := \frac{1-2m}{3(2-m)}$ (which belongs to $(0, 1/6)$).

Then we get in particular,

$$\lim_{t \to +\infty} \rho(t)[U] \ = \ \min_{M} U$$

or equivalently, for any $\epsilon > 0$,

$$\lim_{t \to +\infty} \rho(t)[U \geqslant \min_{M} U + \epsilon] \ = \ 0$$

i.e. the concentration of $\rho(t)$ around $\mathcal{M}$.

Plan

A non-linear Markov process

Let us come back to (4). A corresponding non-linear diffusion is a continuous stochastic process $(X(t))_{t\geqslant 0}$ satisfying

$$\forall\ t \geqslant 0, \quad \left\{ \begin{array}{rcl} dX(t) & = & \sqrt{2\alpha(\rho_t(X(t))}dB(t) - \beta_t \nabla U(X(t))\, dt \\ \mathcal{L}(X(t)) & = & \rho_t\, d\ell \end{array} \right.$$

say on the torus $\mathbb{T}^D$ of dimension $D \geqslant 1$. This formulation is non-linear, due to the presence of ρ_t, the density of the law of $X(t)$.

The existence and uniqueness of such processes is not obvious, but some results can be found in the literature.
The direct sampling of $(X(t))_{t\geqslant 0}$ remains problematic.

Regularization

The situation becomes easier through some regularization procedure. Consider $K : \mathbb{T}^D \to \mathbb{R}_+$ a smooth function with a support localized in a small ball and satisfying $\int K \, d\ell = 1$. For any probability density ρ on M and a bandwidth parameter $h \in (0, 1)$, set

$$\forall\, x \in M, \qquad \rho_h(x) \;:=\; h^{-D} \int_M K((x-y)/h)\, \rho(y) \ell(dy)$$

We can replace the previous non-linear sde by

$$\forall\, t \geqslant 0, \qquad dX(t) \;=\; \sqrt{2\alpha(\rho_{h,t}(X(t))}\, dB(t) - \beta_t \nabla U(X(t))\, dt$$

whose non-linearity is less radical and thus simpler to investigate via classical mean field theory, even if h was to depend on time and going to zero in large time.

Particles

The previous evolution can be approximated by interacting particles systems. Consider a system of N particles, $X_1, X_2, ..., X_N$ whose joint evolution is described by the stochastic differential equations,

$$\forall\ n \in [\![N]\!], \quad dX_n(t) \quad = \quad -\beta_t \nabla U(X_n(t)) + \sqrt{\alpha(\rho_{N,h,t}(X_n(t)))}\, dB_n(t)$$

where the $(B_n(t))_{t \geqslant 0}$, for $n \in [\![N]\!]$, are independent Brownian motions of dimension d, and where

$$\forall\ x \in M, \qquad \rho_{N,h,t}(x) \quad := \quad \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} K\left(\frac{x - X_n(t)}{h}\right)$$

This quantity "counts" the number of particles which are close to x in the scale $h \ll 1$.

Hope

For large N and small h, $\rho_{N,h,t}$ should be close to ρ_t, with the advantage to be samplable. Ideally, both N and h should also depend on time, and respectively going to infinity and zero (with certain rates to be determined...). In particular new particles are born as N increases. The corresponding stochastic algorithm should provide a new global swarm optimization procedure. It seems to perform better than simulated annealing in preliminary numerical experiments provided by Lénaïc Chizat comparing φ_1 avec $\varphi_{1/2,2}$. In particular, compare at the end of the simulations the concentrations of the particules at the bottom of the wells: for $\varphi_{1/2,2}$ they are in better positions to go towards the well containing the global minimum.

Plan

References on simulated annealing

- R. Azencott, editor. *Simulated annealing. Parallelization techniques*. Chichester: John Wiley & Sons Ltd., 1992.
- B. Hajek. Cooling schedules for optimal annealing. *Math. Oper. Res.*, 13(2):311–329, 1988.
- 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.
- L. Miclo. Recuit simulé sur $\mathbf{R}^n$. Étude de l'évolution de l'énergie libre. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(2), p. 235-266, 1992.

References on Wasserstein space


L. Ambrosio, N. Gigli and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.


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


F. Santambrogio. *Optimal transport for applied mathematicians*. volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015.


C. Villani. *Optimal transport, Old and new.* volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

References on swarm-type algorithms

- J. Bolte, L. Miclo, and S. Villeneuve. Swarm gradient dynamics for global optimization: the density case. arXiv preprint, 2022.
- M. Dorigo and C. Blum. Ant Colony optimization theory: a survey. *Theoretical Computational Science.*, 344, p. 243–278, 2005.
- D.B. Fogel. Evolutionary computation towards a new philosophy of machine intelligence. IEEE Press, NJ, Second Ed., 2000.
- J. Kennedy and R. Eberhart. Particle Swarm optimization. *In Proceedings of Int. Conf. on neural networks.*, 4:1942–1946, 1995.