On the convergence of stochastic algorithms

Laurent Miclo
(based on works with Marc Arnaudon)

Plan

- The problem of finding Fréchet means
- Simulated annealing
- Functional inequalities
- A stochastic algorithm finding Fréchet means
- Discrepancies between probabilities measures

- Proof of convergence

A priori hope: each of the 1h30 lectures will consist in approximatively two of the above consecutive points.

1- Fréchet means

Consider (M, d) a compact metric space, endowed with a probability measure ν.

Fix $p \in [1, +\infty[$ and define the mapping

$$U : M \longrightarrow \mathbb{R}$$
$$x \longmapsto \int d^p(x, y)\, \nu(dy)$$

Denote M_p the set of global minima of U.

An element x of M_p is

called a p-mean associated to ν.

For $p=2$, π is simply called a mean, while for $p=1$, π is a median.

In the above definition the compactness assumption ensures that M_p is not empty, but the notion of p-means can be extended to general metric probability

spaces, under the condition that ν admits a p-moment:

$\exists x \in M$ (equivalently $\forall x \in M$):

$$\int d^p(x, y)\, \nu(dy) < +\infty$$

When $M = \mathbb{R}^k$, with $k \in \mathbb{N}$, endowed with the Euclidean distance, one recovers the usual notion of mean, or median (for $k = 1$).

But in general the p-mean is not unique: consider for

instance the Haar measure ν on the circle $\mathbb{T}$ endowed with the Riemannian distance inherited from its usual embedding in $\mathbb{R}^2$:

$$\forall p \geq 1, \quad M_p = \mathbb{T}$$

A general problem is to find M_p, from a sequence $(Y_n)_{n \in \mathbb{N}}$ of i.i.d. random variables distributed according to ν (in practice, the sample is

rather finite but large...).

In these lectures, we will only consider the Riemannian setting, where numerous applications are already taking place. For instance some infinite dimensional extensions of this framework are important in image processing: the distance between two images is obtained by associating some weights to natural infinitesimal transformations

enabling to pass from one to the other.

Other contexts are interesting too, e.g. weighted graphs: what is the mean size of a network, where the pre-distance between two sites is inversely proportional to the traffic between them?

So to end this introduction, let us recall the necessary notions from Riemannian geometry.

By definition, we are given on each tangent space of the compact manifold M a scalar product $\langle \cdot, \cdot \rangle$. It enables to define the gradient operator from $C^\infty(M) \to C^\infty(TM)$ by

$$\forall f \in C^\infty(M), \forall x \in M, \forall v \in T_x M,$$

$$df_x(v) = \langle \nabla f, v \rangle_x$$

There is a natural measure μ associated to the Riemannian structure: in local coordinates

if $\langle \cdot, \cdot \rangle = \sum_{k,l} g_{k,l}(x)\, dx^k dx^l$,

$$\mu(dx) = \sqrt{\det(g(x))}\, dx^1 \dots dx^N$$

(with $N = \dim(M)$).

In the compact setting μ has finite weight, so we renormalize it to get a probability measure. The divergence operator is defined by duality on vector fields b via

$$\forall f \in C^\infty(M), \int \langle b, \nabla f \rangle \, d\mu = -\int \operatorname{div}(b) f \, d\mu$$

(in local coordinates, $\operatorname{div}(b)$

$$= \frac{1}{\sqrt{\det(g)}} \sum_k \partial_k \left(\sqrt{\det(g)}\, b^k \right) \Big).$$

The Laplace-Beltrami operator is then given by

$$\Delta \cdot = \operatorname{div}(\nabla \cdot)$$

$$\left(= \frac{1}{\sqrt{\det(g)}} \sum_{k,l} \partial_k g^{k,l} \sqrt{\det(g)}\, \partial_l \cdot \right).$$

which leads to the important integration by part formula:

$\forall f, g \in C^\infty(M)$,

$$\int f \Delta g \, d\mu = \int g \Delta f \, d\mu$$

$$= -\int \langle \nabla f, \nabla g \rangle \, d\mu$$

The associated distance is given by

$$\forall x, y \in M$$

$$d(x,y) := \min_{\gamma : x \to y} \int_0^T \underbrace{\sqrt{\langle \dot\gamma_t, \dot\gamma_t \rangle}}_{=: \|\dot\gamma_t\|} \, dt$$

where the minimum is over all smooth paths $\gamma : [0,T] \to M$ such that $\gamma(0) = x$, $\gamma(T) = y$.

A minimum path γ with $\|\dot\gamma_t\| \equiv 1$ is called a geodesic

with unitary speed. For x, y sufficiently close, it is unique and will be denoted $\gamma_{x,y}$. It satisfies the Euler-Lagrange equation which enables to define for all times $t \in \mathbb{R}$.

2- Simulated annealing

On a compact Riemannian manifold M, let be given U a smooth function. We would like to find the set M_* of global minima of U.

The corresponding simulated annealing algorithm is a time-inhomogeneous Markov process $(X_t)_{t \geq 0}$, described in terms of its generators

$(L_{\beta_t})_{t \geq 0}$:

$\forall \beta \geq 0$,

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

and $\mathbb{R}_+ \ni t \longmapsto \beta_t \in \mathbb{R}_+$

is a smooth evolution of the inverse of the temperature.

to be determined so that

for any neighborhood $\mathcal{N}$ of

$\mathcal{U}$, we have

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

(so that to find with a good chance a point close to $\mathcal{M}$, it is sufficient to pick X_t for $t \geq 0$ large enough).

Rigorously, $X := (X_t)_{t \geq 0}$ is such that for any function $f \in C^\infty(M)$, the process M^f defined by

$$\forall t \geq 0, \quad M_t^f := f(X_t) - f(X_0) - \int_0^t L_{p_s} f(X_s)\, ds$$

is a martingale in the filtration generated by X.

Due to our regularity assumptions, once $\mathcal{L}(X_0)$ (the law of X_0)

is given, the above martingale problem uniquely determines $\mathcal{L}(X)$ on $C(\mathbb{R}_+, M)$.

There are probabilistic constructions of X which give a meaning to the description

$$dX_t = \sqrt{2}\,\sigma(X_t)\,dB_t - \beta_t \nabla U(X_t)\,dt$$

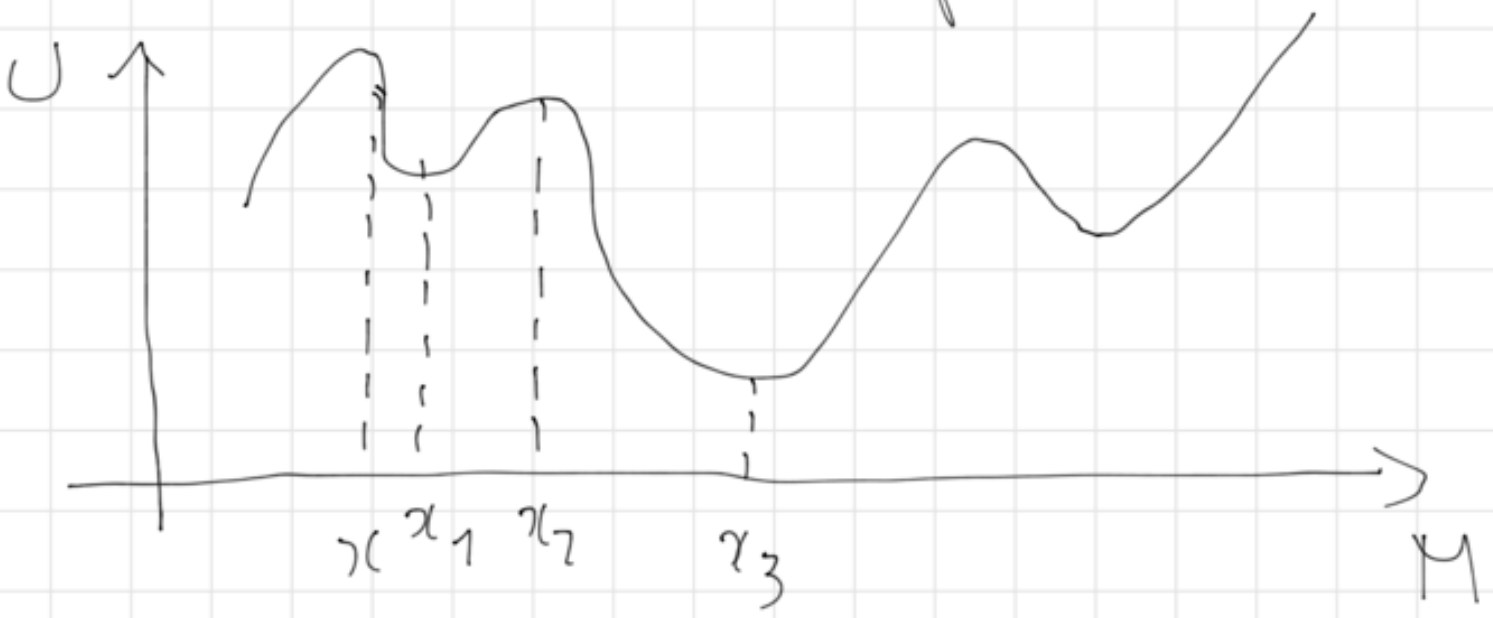
where for any $x \in M$

$$\sigma(x) : \mathbb{R}^N \to T_x M$$

is such that $\sigma(x)\sigma^*(x) = Id$ (and where the matrix $\sigma(x)$ depends smoothly on x).

In the above formula, $(B_t)_{t \geq 0}$ is a standard Brownian motion in $\mathbb{R}^N$.

We thus have the picture



At x, X_t "feels" the influence of $-\beta_t \nabla U(X_t)$, which pushes it toward the local minimum x_1, as well as the perturbation

by the Brownian motion, which will help it to escape from the well whose bottom is x_1, to pass through x_2 in direction of the global minimum x_3.

To prevent nice evolutions $\beta := (\beta_t)_{t \geq 0}$, we need to introduce a critical constant $b(U)$.

If $\rho : [0,1] \to M$ is a continuous path, we denote

its elevation by

$$U(p) = \max_{t \in [0,1]} U(p(t))$$

The minimal elevation between two points $x, y \in M$ is

$$U(x,y) = \min_{p \in \mathcal{P}_{x,y}} U(p)$$

$\mathcal{P}_{x,y}$ = set of continuous paths going from x to y.

Then we consider

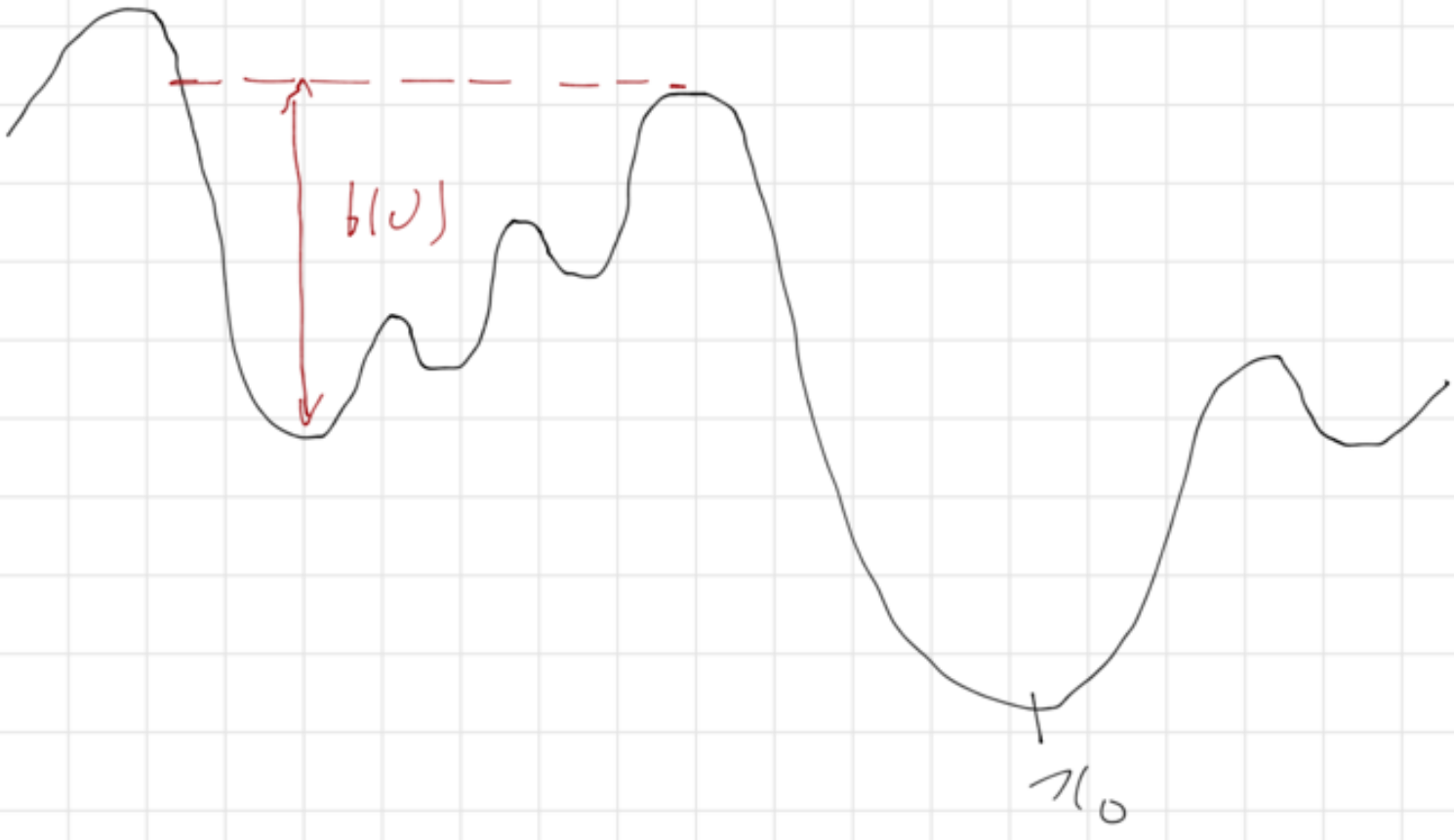
$$b(U) := \max_{x,y \in M} U(x,y) - U(x) - U(y) + \min_M U$$

Fix $x_0 \in M$, this constant

is also equal to

$$b(U) = \max_{y \in M} U(x_0, y) - U(y)$$

namely, $b(U)$ is the highest height of a well not containing x_0:



For $\lambda \geq 1$ and $b > 0$, consider the evolution of β given by

$$\forall t \geq 0, \quad \beta_t = b^{-1} \ln(1 + t)$$

Holley, Kusuoka and Stroock [1989] have proven the following result:

Theorem

If $b > b(U)$, then $\forall \mathcal{L}(X_0)$,

$$\lim_{t \to +\infty} \| \mathcal{L}(X_t) - \mu_{\beta_t} \|_{TV} = 0$$

If $b < b(U)$, then $\exists \mathcal{L}(X_0)$:

$$\lim_{t \to +\infty} \| \mathcal{L}(X_t) - \mu_{\beta_t} \|_{TV} \neq 0$$

where for any $\beta \geq 0$, μ_β is the Gibbs measure associated to the potential U and to the temperature β:

$$\mu_\beta(dx) = \frac{e^{-\beta U(x)}}{Z_\beta} \, \mu(dx)$$

$Z_\beta \rightarrow$ normalizing constant

Note that for $\beta \geq 0$ large, μ_β concentrates around $\mathcal{W}$:

For any neighborhood $\mathcal{N}$ of $\mathcal{W}$,

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

It follows that for the evolution of β corresponding to $\Lambda \geq 1$ and $b > b(U)$, our goal (1) is fulfilled (recall the definition of the total variation norm between two probability measures m and μ:

$$\|m - \mu\|_{TV} = 2 \sup_{A \subset M \text{ Borelian}} m(A) - \mu(A)$$

$$= \sup_{f \in C^0(M),\ \|f\|_\infty \leq 1} m[f] - \mu[f]$$

)

Note that μ_β is reversible for L_β: $\forall f, g \in C^\infty(M)$,

$$\mu_\beta[f L_\beta g] = \mu_\beta[g L_\beta f]$$

Indeed, we have

$$L_\beta \cdot = e^{\beta U} \operatorname{div}(e^{-\beta U} \nabla \cdot)$$

so that

$$\mu_\beta[f L_\beta g] = \frac{1}{Z_\beta} \mu[f \operatorname{div}(e^{-\beta U} \nabla g)]$$

$$= -\frac{1}{Z_\beta} \mu[\langle \nabla f, e^{-\beta U} \nabla g \rangle]$$

$$= -\mu_\beta[\langle \nabla f, \nabla g \rangle]$$

In particular μ_β is invariant for L_β, thus at any time $t \geq 0$, $\mathcal{L}(X_t)$ tends instantaneously to get closer to μ_{β_t}. The above theorem states that if $(\beta_t)_{t \geq 0}$ is changing sufficiently slowly ($b > b(U)$), then $\mathcal{L}(X_t)$ succeeds to follow its target μ_{β_t} and to get closer and closer to it in large times.

Remarks:

a) For $b > b(U)$, a.s. convergence

doesn't hold for X: it can be shown that the recurrence set of X is a.s. the connected component of $\{x : U(x) \leq \min_H U + b\}$ containing $\mathcal{M}$, even if most of the time X_t is close to $\mathcal{M}$ for large $t \geq 0$.

b) The constant $b(U)$ is not critical for (1), it is the constant $\tilde{b}(U)$ defined by

$$\tilde{b}(U) := \max_{y \in H} \min_{x_0 \in \mathcal{M}} (U(x_0, y) - U(y))$$

A simple proof of the first part of the theorem consists in studying the evolution of a discrepancy between $m_t := \mathcal{L}(X_t)$ and $\mu_{\beta t}$, which is easier to differentiate than the total variation (absolute value is rather troublesome to manipulate).

In the above situation, the most appropriate choice is the relative entropy $\mathrm{Ent}(m_t, \mu_{\beta t})$, but to avoid the investigation

of logarithmic Sobolev inequalities,

let us rather consider the functional

$\forall t > 0,$

$$I_t := \int \left(\frac{d\nu_t}{d\mu_{\beta_t}} - 1 \right)^2 d\mu_{\beta_t}$$

$\frac{d\nu_t}{d\mu_{\beta_t}}$ ↑ Radon-Nikodym density,

well-defined for $t > 0$ by

ellipticity of the generators

By Cauchy-Schwarz inequality,

$$\| \nu_t - \mu_{\beta_t} \|_{TV} = \int \left| \frac{d\nu_t}{d\mu_{\beta_t}} - 1 \right| d\mu_{\beta_t}$$

$$\leq \sqrt{I_t}$$

it is thus sufficient to show that

$$\lim_{t \to +\infty} I_t = 0$$

To study the evolution of $t \longmapsto I_t$, we must know how to differentiate m_t with respect to time. By the martingale problem, we have for any $f \in \mathcal{C}_b^\infty(\mathbb{T}^1)$,

$$\partial_t \mathbb{E}[f(X_t)] = \mathbb{E}[L_{\beta_t}[f](X_t)]$$

namely, in the weak sense

$$\partial_t \langle f, m_t \rangle = \langle L_{\beta_t}[f], m_t \rangle$$

where the probability measure

m_t and its density $\frac{dm_t}{d\mu}$ are written in the same manner m_t (for $t>0$, $m_t \in C^\infty(M)$).

It will be convenient to denote

$$f_t := \frac{dm_t}{d\mu_{\beta t}}, \text{ for } t>0.$$

We can now differentiate with respect to time: for $t>0$,

$$I_t' =$$

$$2\int (f_t - 1)\left(\frac{\partial_t m_t}{\mu_{\beta t}} - f_t \partial_t \ln \mu_{\beta t}\right) d\mu_{\beta t}$$

$$+ \int (f_t - 1)^2 \partial_t \ln \mu_{\beta t}\, d\mu_{\beta t}$$

$$= 2 \int (f_t - 1) \partial_t \ln \mu_{\beta_t} \, d\mu$$

$$- \iint \left((f_t - 1)^2 + 2 (f_t - 1) \right) \partial_t \ln \mu_{\beta_t} \, d\mu_{\beta_t}$$

$$\leq 2 \int L_{\beta_t} (f_t - 1) \, f_t \, d\mu_{\beta_t}$$

$$+ \| \partial_t \ln \mu_{\beta_t} \|_\infty \left(I_t + 2 \sqrt{I_t} \right)$$

$$= -2 \int \| \nabla f_t \|^2 \, d\mu_{\beta_t}$$

$$+ \| \partial_t \ln \mu_{\beta_t} \|_\infty \left(I_t + 2 \sqrt{I_t} \right)$$

The last term is easy to evaluate:

$$\| \partial_t \ln \mu_{\beta_t} \|_\infty = \| \partial_\beta \ln \mu_\beta \|_\infty \, |\beta_t'|$$

and since for $x \in M$

$$\ln \mu_\beta(x) = -\beta U(x) + \ln\left(\int e^{-\beta U} d\mu\right)$$

$$\|\partial_\beta \ln \mu_\beta\|_\infty \leq \operatorname{osc}(U)$$

$$:= \max_M U - \min_M U$$

Concerning the first term, we need a spectral gap estimate due to Holley, Kusuoka and Stroock [1987]:

Proposition

There exists a constant

C_M depending only on M, such that $\forall \beta \geq 0$, $\forall f \in C^\infty(M)$,

$$\mu_\beta[(f-1)^2] \leq c(\beta)\, \mu_\beta[\|\nabla f\|^2]$$

with

$$c(\beta) = C_M (1 \vee (\beta \|\nabla U\|_\infty))^{5N-2} \exp(6(U)\beta)$$

We will recall the proof of this Poincaré inequality in next section. For the time being let $J_t = \sqrt{I_t}$, we get for $t > 0$,

$$J_t' \leq -c_t J_t + K_t$$

with

$$c_t := c(\beta_t) - K_b/2$$

$$K_t := 2\|U\|_\infty |\beta_t'|$$

By Gronwall's lemma, we conclude that

$$\lim_{t \to +\infty} J_t = 0$$

if $$\int^{+\infty} c_t \, dt = +\infty$$

$$\lim_{t \to +\infty} \frac{K_t}{c_t} = 0$$

This is satisfied if $b > b(U)$, because then K_t is of order

$\frac{1}{t}$ and c_t of order $\frac{1}{t^{b/(1+b)}}$

(up to a logarithmic correction).

3- functional inequalities

To avoid technicalities and to concentrate on ideas, we will deduce the wanted Poincaré inequality in the finite state space setting.

The underlying path approach is also due to Holley and Stroock [1988].

M is now a finite set endowed with a Markovian generator matrix $L := (L(x,y))_{x,y \in M}$

Satisfying

$$\forall x \neq y, \quad L(x,y) \geq 0$$

$$\forall x, \quad \sum_y L(x,y) = 0$$

We assume irreducibility:

$\forall x, y: \exists \; x = x_0, x_1, \dots, x_n = y$

with $L(x_i, x_{i+1}) > 0 \quad \forall 0 \leq i < n$ (?)

and reversibility: there exists μ probability measure such that

$$\forall x, y, \quad \mu(x) L(x,y) = \mu(y) L(y,x)$$

The operator L plays the role of the Laplacian.

Let U be a function on M and $\beta \geq 0$. The traditional Metropolis perturbation of L admitting the corresponding Gibbs measure ($\mu_\beta(x) \propto e^{-\beta U(x)}$ $\mu(x)$) for reversible probability is defined via

$$\forall x \neq y, \quad L_\beta(x,y) = e^{-\beta (U(y) - U(x))_+} L(x,y)$$

The associated Dirichlet form is given by

$$\forall f \in \mathcal{F}(M) \quad | = \text{set of functions on } M),$$

$$\mathcal{E}_\beta(f) = -\mu_\beta(f L_\beta f)$$

$$= \frac{1}{2}\sum_{x,y} \mu_\beta(x) L_\beta(x,y) \left(f(y) - f(x)\right)^2$$

$$e^{-\beta(U(x) \vee U(y))} \mu(x) L(x,y)$$

This replaces $\mu_\beta[\|\nabla f\|^2]$

in the continuous setting.

We would like to estimate the best constant $c(\beta)$ such that $\forall f \in \mathcal{F}(M)$

$$\underbrace{\mathrm{Var}(f, \mu_\beta)}_{\mu_\beta[|f - \mu_\beta(f)|^2]} \leq c(\beta)\, \mathcal{E}_\beta(f)$$

Proposition

There exists a constant $C_L > 0$ such that

$$\forall \beta \geq 0, \quad c(\beta) \leq C_L \, e^{\beta b(U)}$$

where $b(U)$ is defined as in the continuous setting, but with the continuous paths replaced by discrete paths satisfying (2)

There is a matching lower bound, but we will not use it.

For the proof, start with the alternative expression for the variance: $\forall f \in \tilde{\mathcal{F}}$,

$$\mathrm{Var}(f, \mu_P) = \frac{1}{2} \sum_{x,y} (f(y) - f(x))^2 \mu_P(x) \mu_R(y)$$

For given $x \neq y$, let $\underset{x,y}{p} = (x = x_0, x_1, \ldots, x_n = y)$ be an admissible path with

$$U(\underset{x,y}{p}) = U(x, y)$$

Then we have

$$(f(y) - f(x))^2 = (f(x_n) - f(x_{n-1}) + f(x_{n-1}) - f(x_{n-2}) + \cdots - f(x_0))^2$$

$$\leq n \sum_{e \in P_{x,y}} (df(e))^2$$

where $P_{x,y}$ is seen as the set of edges $\{(x_0, x_1), (x_1, x_2), \ldots (x_{n-1}, x_n)\}$ and where

$df(e) = f(x'') - f(x')$ if $e = (x', x'')$.

Denote similarly

$$\mu_\beta L_\beta(e) = \mu_\beta(x', L_\beta(x', x''))$$

Let N be the cardinality of $\mathcal{H}$. We have

$$\mathrm{Var}(f, \mu_\beta) \leq \frac{N}{2} \sum_e (df(e))^2 \sum_{\substack{x,y \\ e \in P_{x,y}}} \mu_\beta(x) \mu_\beta(y)$$

$$= \frac{N}{2} \sum_e (df(e))^2 \mu_\beta L_\beta(e) \, A_\beta(e)$$

where

$$A_\beta(e) := \frac{1}{\mu_\beta L_\beta(e)} \sum_{\substack{x,y \\ e \in P_{x,y}}} \mu_\beta(x) \mu_\beta(y)$$

$$\cdot \frac{e^{\beta(U(x') \vee U(x''))}}{Z_\beta \mu_\beta(x') L_\beta(x', x'')} \sum_{\substack{x',y' \\ e \in P_{x',y'}}} e^{-\beta(U(x') + U(y'))} \mu_\beta(x') \mu_\beta(y')$$

with $e = (x', x'')$.

Using that

$$Z_\beta \asymp e^{-\beta \min_M U}$$

up to upper and lower bounds independent of β

we get that

$$\max_{e \in P_{n,g}} A_\beta(e) \asymp e^{\beta(U(\pi,g) - U(\pi) - U(g) + \min_\pi U)}$$

so

$$\max_e A_\beta(e) \asymp e^{\beta b(U)}$$

which is what we wanted.

Note: it is sometimes more convenient to work with the relative entropy

$$Ent(m, \lambda) = \begin{cases} \int \ln \frac{dm}{d\mu} \, dm, & \text{if } m \ll \mu \\ +\infty, & \text{otherwise} \end{cases}$$

than with the quantity considered in the previous section.

In particular, we have Pinsker's inequality:

$$\|\mu - \nu\|_{TV} \leq \sqrt{2} \sqrt{\mathrm{Ent}(\mu, \nu)}$$

For the previous approach to work, the Poincaré inequality must be replaced by a logarithmic Sobolev inequality: $\forall f \in \mathcal{F}, f \geq 0, f \not\equiv 0$

$$\mathrm{Ent}\left(\frac{f}{\mu_\beta(f)} \mu_\beta, \mu_\beta\right) \leq \tilde{C}(\beta) \mathcal{E}_\beta(f)$$

It can be shown that the best constant $\tilde{C}(\beta)$ in this bound has at the exponential level a behavior similar to $r(\beta)$:

$$\exists \tilde{C}_L > 0 : \forall \beta \geq 0$$

$$\tilde{C}(\beta) \leq \tilde{C}_L (1+\beta) e^{b(U)\beta}$$

This is interesting on very large spaces, where the use of specific quantities (renormalisation by the number of sites, e.g.) is required.

4- Another stochastic algorithm

We come back to the problem of finding the Fréchet mean of a probability measure ν on a compact Riemannian manifold M. We have at our disposal a sequence $(Y_n)_{n\in\mathbb{N}}$ of r.v. distributed according to ν.

We want to use them to construct an algorithm $(X_t)_{t\geq 0}$ finding μ: For all neighborhood

$$\mu = \mu_2$$

N of ω,

$$\lim_{t \to +\infty} \mathbb{P}[X_t \in N] = 1 \quad (3)$$

There is a proposition: in addition to the inverse temperature schedule $(\beta_t)_{t \geq 0}$, let be given $(\alpha_t)_{t \geq 0}$ an homogenization schedule, tuning the speed of use of the data.

More precisely, let $(N_t)_{t \geq 0}$ be a usual Poisson process (interjumps : independent

exponential variables of parameter 1).

Consider the time changed process $N^{(\alpha)}$:

$\forall t \geq 0$, $N_t^{(\alpha)} := N_{\int_0^t 1/\alpha_s \, ds}$

We assume that $\alpha_t \underset{t \to +\infty}{\searrow} 0$

so the jumps of $N^{(\alpha)}$ occur more and more often. Call them $0 < T_1 < T_2 < \cdots < T_n < \cdots$,

we have that $T_n \underset{n \to +\infty}{\longrightarrow} +\infty$

Between these times, X

evolves as a Brownian motion,

namely its generator is $\frac{\Delta}{2}$.

At time T_n, X jumps

from $X_{T_n^-}$ to $\gamma_{X_{T_n^-}, Y_n}(\delta_n)$

with

$$\delta_n = \alpha_{T_n} \beta_{T_n} d(X_{T_n^-}, Y_n)$$

Note that by construction

$\mathcal{L}(X_{T_n^-}) \ll \mu$, so the

event that $X_{T_n^-}$ is in the

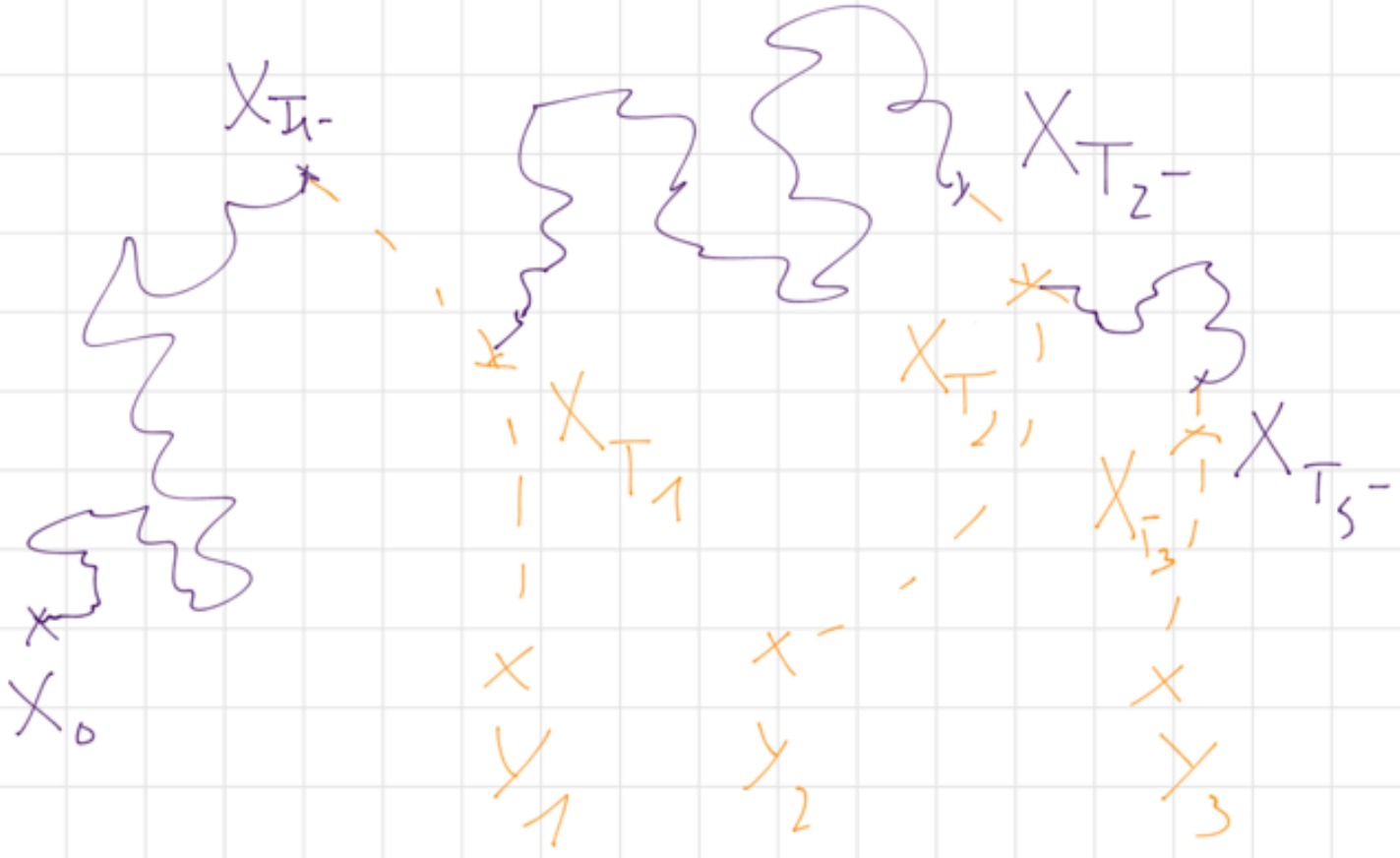
sub-locus of Y_n is negligeable.

Furthermore, α and β will

satisfy

$$\lim_{k-1+\alpha} \alpha_k \beta_k = 0$$

thus X_{T_k} will end up being quite close to $X_{T_{k-}}$: as time goes on, there will be a lot of small jumps:



Conjecture

Assume that ν admits a density (wrt μ) Hölder continuous of exponent $a \in (0,1]$. Then there exists $\bar{a} > 0$ depending on a and M such that (3) is fulfilled if $\forall k \geq 0$,

$$\begin{cases} \alpha_k = (1+k)^{-\frac{1}{\bar{a}}} \\ \beta_k = b^{-1} \ln(1+k) \end{cases}$$

with $b > b(U)$

This result is true for the circle $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$:

Theorem

The above conjecture is true for $M = \Pi$ if $\tilde{a} = a$.

More generally, it is true for p-means, if we take

$$s_n = \frac{p}{2} \alpha_{T_n} \beta_{T_n} d^{p-1}(X_{T_n - 1}, X_n)$$

and

$$\tilde{a} = \begin{cases} 1 & \text{, if } p = 1 \text{ or } p \geq 2 \\ \min(a, p-1) & \text{, if } p \in (1,2) \end{cases}$$

There is also a version valid for general probability

measure ν: one needs to add some time-inhomogeneous noise to the sequence $(Y_n)_{n\in\mathbb{N}}$.

More precisely, for $x \in M$ and $K > 0$, consider on $T_x M$ the probability measure $\widetilde{K}_{x,K}$ whose density with respect to the Lebesgue measure is proportional to $(1 - K\|v\|)_+$.

Denote by $K_{x,K}$ the image of $\widetilde{K}_{x,K}$ by the exponential mapping.

Construct the process $(Z_t)_{t\geq 0}$ as X but by replacing Y_n

by a point sampled according to $K_{y_n, x_n}(\cdot)$, where $(x_t)_{t \geq 0}$ is another schedule to tune.

If $M = \Pi$ and $p = 2$, we have:

Theorem

$$\lim_{t \to +\infty} \mathbb{P}[Z_t \in \mathcal{N}] = 1$$

For any neighborhood $\mathcal{N}$ of $\mathcal{U}$ if

$$\forall t \geq 0 \begin{cases} \alpha_t = (1+t)^{-(2k+1)} \\ \beta_t = b^{-1} \ln(1+t) \\ x_t = (1+t)^k \end{cases}$$

with $b > b(U)$, $k > 0$.

How to choose k?

The larger k is, the closer ν will be to its approximation by its transport through the kernel $K_{\cdot, x_t}(\cdot)$ and the above convergence will be more precise. But the faster the algorithm uses the data $(Y_n)_{n \in \mathbb{N}}$. So in practice a trade-off has to be made between the number of r.v. Y_n at our disposal and the quality of the approximation of μ.

Heuristic of the convergence in the first theorem.

For $\alpha > 0$, $\beta \geq 0$ consider the generator defined by

$$\forall f \in C^\infty(M),\ \forall x \in M,$$

$$L_{\alpha,\beta}[f](x) = \frac{\Delta}{2} f(x) + \frac{1}{\alpha} \int f(\gamma_{x,y}^{1}(\Delta)) - f(x) \, \nu(dy)$$

with $\Delta = \beta \alpha d(x,y)$.

At any time $t \geq 0$, L_{α_t, β_t} is the instantaneous generator of X.

We have, for fixed x, y, β,

$$\lim_{\alpha \to 0_+} f(\gamma_{x,y}(1)) - f(x)$$

$$= \beta \, d(x,y) \langle \nabla f(x), \dot{\gamma}_{x,y}(0) \rangle$$

for any $f \in C^\infty(M)$ and where

$$\dot{\gamma}_{x,y}(0) = \frac{d}{dt} \gamma_{x,y}(t) \Big|_{t=0} \in T_x M$$

Thus

$$\lim_{\alpha \to 0_+} L_{\alpha,\beta}[f](x) = \frac{\Delta}{2} f(x) + \frac{\beta}{2} \langle F, \nabla f \rangle_x$$

where

$$F(x) = 2 \int d(x,y) \, \dot{\gamma}_{x,y}(0) \, \nu(dy)$$

$$= - \int \nabla_x d^2(x,y) \, \nu(dy)$$

$= -\nabla U(x)$

(still with $p = 2$).

We recover the generator of the simulated annealing algorithm associated to U, explaining why we expect that X will find $\mathcal{U}$ in large time.

But more rigorous computations are needed to justify the choice of $(\alpha_t)_{t \geq 0}$. The same heuristic enables to understand

the convergence of $(Z_t)_{t \geq 0}$.

One drawback of these algorithms is the computations of the geodesic paths $\gamma_{x,y}$. In general it is not easy, but this task becomes simpler for symmetrical spaces. For instance if M is a sphere (of any dimension ≥ 1)

$$\gamma_{x,y}(\Delta) = \cos(\Delta)\, x + \sin(\Delta)\, \frac{y - \langle y, x \rangle x}{\sqrt{1 - \langle y, x \rangle^2}}$$

(for $|\langle y, x \rangle| \neq 1$).

Furthermore, still on the sphere, one does not need to entirely simulate trajectories of the Brownian motion: one can use the 1-dimensional Jacobi diffusion obtained by projection on the radius as well as a simple uniform sampling on subspheres. Similar remarks are valid for hori. Let us illustrate that with Π: to simulate $X_{T_1^-}$, one sample

G a standard Gaussian distribution and reduce modulo 2π $X_0 + \sqrt{\gamma_1}\, G$.

The above algorithm can be extended to treat more general cases, but at the expence of difficulties of implementation.

More precisely, consider $\rho : M \times M \to \mathbb{R}$ a continuous function and

$$U : r \longmapsto \int \rho(r, y)\, \nu(dy)$$

In particular, ρ can be d or d^p, with $p > 0$, or the

length associated to a Finsler metric on M (still a compact Riemannian manifold).

Denote $(p(\delta, x, y))_{\delta>0,\, x,y \in M}$ the heat kernel on M. We will use it to regularize the left argument of c.

So for $\delta > 0$, consider

$$p_\delta(x,y) := \int_M p(\delta, x, z)\, p(z, y)\, \mu(dz)$$

and

$$U_\delta : x \longmapsto \int_M p_\delta(x,y)\, \nu(dy)$$

In addition to β and α, let approximations be given an schedule $(\delta_t)_{t \geq 0}$ with values in $(0, +\infty]$.

A process X is constructed as before, except that at time T_n, the process jumps from $X_{T_n^-}$ to

$$X_{T_n} := \phi_{\delta_{T_n}}(\alpha_{T_n}\beta_{T_n}, X_{T_n^-}, Y_n)$$

where for any $x, y \in M$, $(\phi_s(s, x, y))_{s \geq 0}$ is the flow starting from x and associated

to the vector field

$$M \ni z \longmapsto -\frac{1}{2} \nabla_z \rho_S(\cdot, y)$$

(hoping that we are able to construct this flow in practice).

Theorem

(3) is fulfilled if we take

$$\forall t \geq 0 \quad \begin{cases} \alpha_t := 1/(1+t) \\ \beta_t := b \ln(1+t) \\ \delta_t := 1/\ln(2+t) \end{cases}$$

with $b > b(U)$

In the following considerations, rather consider the Gibbs measure

$$\mu_{\beta,S}(dx) = \frac{e^{-\beta U_S(x)}}{Z_{\beta,S}} \mu(dx)$$

normalizing constant

and traditional estimates on the heat kernel.

5 – Discrepancies between measures

At first thought, we would like to study the evolution of

$$t \mapsto D(\mathcal{L}[X_t], \pi_t)$$

where π_t is the instantaneous invariant measure at time t

and where for any probability measures m and μ

$$D(m, \mu) := \begin{cases} \int \left(\frac{dm}{d\mu} - 1\right)^2 d\mu & \text{if } m \ll \mu \\ +\infty & \text{otherwise} \end{cases}$$

Unfortunately, the invariant measure $\Pi_{\alpha,\beta}$ associated to $L_{\alpha,\beta}$ is difficult to manage because of the combination of diffusion and jump parts. So we will rather investigate the evolution of

$$I : t \mapsto D(\mathcal{L}(X_t), \mu_{\beta^{-1}})$$

where μ_β is the Gibbs measure associated to the potential U and to the temperature β^{-1}.

At least, we have already

seen how to estimate its Poincaré constant.

What is missing is an evaluation of the discrepancy between μ_β and $\Pi_{\alpha,\beta}$, for given $\alpha > 0$, $\beta \geq 0$.

Let $L^*_{\alpha,\beta}$ be the dual operator of $L_{\alpha,\beta}$ in $L^2(\mu_\beta)$.

A crucial object is $L^*_{\alpha,\beta}[\Pi]$, the smaller it is, the closer μ_β is to $\Pi_{\alpha,\beta}$. We have e.g.

$$L^*_{\alpha,\beta}[\Pi] = 0 \iff \mu_\beta = \Pi_{\alpha,\beta}$$

Indeed, just use that

$\forall f \in C^\infty(M)$,

$$\mu_\beta[L_{\alpha,\beta}[f]] = \mu_\beta[L^*_{\alpha,\beta}[\pi] f]$$

A priori $L^*_{\alpha,\beta}[\pi]$ is not out of reach, since it is constructed from $L_{\alpha,\beta}$ and μ_β which are given.

Thus $\forall f \in C^\infty(M)$, we compute via a straightforward integration by parts, that

$$L^*_{\alpha,\beta}[f] = \frac{1}{2} e^{\beta U} \nabla^2 (e^{-\beta U} f) +$$

$$\frac{e^{\beta U}}{\alpha(1-\alpha\beta)} \int_{\text{ball}} \mathcal{O}(y, (1-\alpha\beta)\pi)\, T_{y, \frac{-\alpha\beta}{1-\alpha\beta}} \left[e^{-\beta U} f \right] \nu(dy) - \frac{f}{\alpha}$$

where $T_{y,s}$ is the operator defined by

$$\forall \eta \in \Pi, \quad T_{y,s} f(\eta) = f\left(\gamma_{\eta, y}(s\, d(\eta, y))\right)$$

One has to be a little careful about the domain of $L^{*}_{\alpha,\beta}$, but as soon as ν admits a continuous density, it is the space of functions whose second derivative in the distribution sense belongs to $L^2(\mu)$.

We need next some computations, based on not very exciting expansions, to get:

Proposition:

Assume that the density of ν satisfies

$$\forall \pi, \eta \in \overline{\Pi}, \quad |D(\eta) - D(\pi)| \leq A\, d^a(\pi, \eta)$$

with $a \in (0,1]$ and $A > 0$.

Then for $\beta \geq 1$ and $\alpha \in (0, 1/(2\beta^2))$,

$$\| L^*_{\alpha,\beta} \Pi \|_\infty \leq \underbrace{C(A)}_{\text{constant depending only on } A} \max(\alpha\beta^4, \alpha^a \beta^{1+a})$$

Similar estimates are also valid for $p \neq 2$ (curiously they are more involved for $1 < p < 2$), or in the case of a continuous ρ:

$$\forall\, \beta \geq 1, \forall\, \delta \in (0,1], \forall\, \alpha \in (0, \delta^2/(2\beta^2)),$$

$$\| L^{\Delta}_{\alpha,\beta,\delta}[\Pi] \|_{\infty} \leq C \alpha \beta^4 \delta^{-4}$$

for an appropriate constant $C > 0$.

6 - End of the proof

Recall that for $t > 0$,

$$I_t := D\left(\mathcal{L}(X_t) \,|\, \mu_{\beta_t}\right)$$

Its differentiation starts as in the simulated annealing case:

$$I_t' \leq 2 \int L\alpha_{t,\beta_t}[f_t - 1] f_t \, d\mu_{\beta_t}$$

$$+ 2 \underbrace{\|U\|_0}_{\leq (\operatorname{diam}(M))^2} |\beta_t'| \left(I_t + 2\sqrt{I_t}\right)$$

where $f_t = \dfrac{d\mathcal{L}(X_t)}{d\mu_{\beta_t}}$

To go further, we write

$$\int L_{\alpha_{t+1},\beta_{t+1}}[f_t - 1] f_t \, d\mu_{\beta_{t+1}} =$$

$$\int L_{\alpha_{t+1},\beta_t}[f_t - 1](f_t - 1) \, d\mu_{\beta_t}$$

$$+ \underbrace{\int L_{\alpha_{t+1},\beta_t}[f_t - 1] \, d\mu_{\beta_t}}$$

$$= \int (f_t - 1) L^{*}_{\alpha_{t+1},\beta_t}[1] \, d\mu_{\beta_t}$$

$$\leq \sqrt{I_t} \, \| L^{*}_{\alpha_{t+1},\beta_t}[1] \|_\infty$$

Let us decompose

$$L_{\alpha,\beta} = \underbrace{L_\beta}_{\Delta} + R_{\alpha,\beta}$$

$$\underbrace{\Delta}_{Z} - \beta \langle \nabla U, \nabla \cdot \rangle$$

We have already seen how to deal with the energy

$$\int (f_t - 1) L_{\beta_t} (f_t - 1) \, d\mu_{\beta_t}$$

via Poincaré inequalities at small temperature.

To treat the term

$$\int (f_t - 1) R_{\alpha_t, \beta_t} (f_t - 1) \, d\mu_{\beta_t}$$

more expansions are needed ...

in the end, we can find a constant $C'(A) > 0$ such that it is bounded above by

$$C'(A)\,\alpha_t\,\beta_t^3 \Big(\underbrace{\int (\nabla f_t)^2 \, d\mu_{\beta_t}} + \int (f_t - 1)^2 \, d\mu_{\beta_t} \Big)$$

$$- \int (f_t - 1) L_{\beta_t} [f_t - 1] \, d\mu_{\beta_t}$$

Putting everything together, it appears that $J_t = \sqrt{I_t}$ satisfies

$$J_t' \leq -\eta_t J_t + \xi_t$$

with

$$\eta_t := c_1(A)\left(\beta_t^{-3}\exp(-b(U)\beta_t) - \alpha_t\beta_t^3 - |\beta_t'|\right)$$

$$\xi_t := c_2(A)\left(\alpha_t\beta_t^4 + |\beta_t'|\right)$$

for two constants $c_1(A)$, $c_2(A)$

depending on A.

It follows that if

$$\lim_{t\to+\infty}\beta_t = +\infty$$

$$\int^{+\infty}\beta_t^{-3}e^{-b(U)\beta_t}\,dt = +\infty$$

$$\max\left(\alpha_t\beta_t'', \alpha_t\beta_t^3, |\beta_t'|\right) << e^{-b(U)\beta_t} \quad t\to+\infty$$

then $\lim_{t\to+\infty} J_t = 0$

This is in particular the case for the explicit schedules α and β given in the first theorem of Section 4.