STOCHASTIC ALGORITHMS FOR COMPUTING MEANS OF PROBABILITY MEASURES

MARC ARNAUDON, CLÉMENT DOMBRY, ANTHONY PHAN, AND LE YANG

Abstract. Consider a probability measure μ supported by a regular geodesic ball in a manifold. For any $p \geq 1$ we define a stochastic algorithm which converges almost surely to the p-mean e_p of μ. Assuming furthermore that the functional to minimize is regular around e_p, we prove that a natural renormalization of the inhomogeneous Markov chain converges in law into an inhomogeneous diffusion process. We give an explicit expression of this process, as well as its local characteristic.

1. Introduction

Consider a set of points $\{x_1, \ldots, x_n\}$ in an Euclidean space E with metric d. The geometric barycenter e_2 of this set of points is the unique point minimizing the mean square distance to these points, i.e.

$$e_2 = \arg\min_{x \in E} \frac{1}{n} \sum_{i=1}^{n} d^2(x, x_i).$$

It is equal to the standard mean $e_2 = \frac{1}{n} \sum_{i=1}^{n} x_i$ and is the most common estimator in statistics. However it is sensitive to outliers, and it is natural to replace power 2 by p for some $p \in [1, 2)$. This leads to the definition of p-means: for $p \geq 1$, a minimizer of the functional

$$H_p : \begin{array}{ccc} E & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & \frac{1}{n} \sum_{i=1}^{n} d^p(x, x_i) \end{array}$$

is called a p-mean of the set of points $\{x_1, \ldots, x_n\}$. When $p = 1$, e_1 is the median of the set of points and is very often used in robust statistics. In many applications, p-means with some $p \in (1, 2)$ give the best compromise. It is well known that in dimension 1, the median of a set of real numbers may not be uniquely defined. This is however an exceptional case: the p-mean of a set of points is uniquely defined as soon as $p = 1$ and the points are not aligned or as $p > 1$. In these cases, uniqueness is due to the strict convexity of the functional H_p.

The notion of p-mean is naturally extended to probability measures on Riemannian manifolds. Let μ be a probability measure on a Riemannian manifold M with distance ρ. For any $p \geq 1$, a p-mean of μ is a minimizer of the functional

$$H_p : \begin{array}{ccc} M & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & \int_M \rho^p(x, y) \mu(dy) \end{array} . \tag{1.1}$$

It should be stressed that unlike the Euclidean case, the functional H_p may not be convex (if $p \geq 2$) and the p-mean may not be uniquely defined. In the case $p = 2$, we obtain the so-called Riemannian barycenter or Karcher mean of the probability measure μ. This has been extensively studied, see e.g. [8], [10], [11], [5], [18],

[2], where questions of existence, uniqueness, stability, relation with martingales in manifolds, behaviour when measures are pushed by stochastic flows have been considered. In the general case $p \geq 1$, Afsari [1] proved existence and uniqueness of p-means on "small" geodesic balls. More precisely, let $\mathrm{inj}(M)$ be the injectivity radius of M and $\alpha^2 > 0$ an upper bound for the sectional curvatures in M. Existence and uniqueness of the p-mean in ensured as soon as the support of the probability measure μ is contained in a convex compact K_μ of a geodesic ball $B(a, r)$ with radius

$$r < r_{\alpha,p} \quad \text{with} \quad r_{\alpha,p} = \begin{cases} \frac{1}{2}\min\left\{\mathrm{inj}(M), \frac{\pi}{2\alpha}\right\} & \text{if } p \in [1,2) \\ \frac{1}{2}\min\left\{\mathrm{inj}(M), \frac{\pi}{\alpha}\right\} & \text{if } p \in [2,\infty) \end{cases}. \tag{1.2}$$

The case $p \geq 2$ gives rise to additional difficulties since the functional H_p to minimize is not necessarily convex any more, due to the fact that we can have $r > \frac{\pi}{4\alpha}$.

Provided existence and uniqueness of the p-mean, the question of its practical determination and computation arises naturally. In the Euclidean setting and when $p = 1$, the problem of finding the median e_1 of a set of points is known as the Fermat-Weber problem and numerous algorithms have been designed to solve it. A first one was proposed by Weiszfeld in [21] and was then extended by Fletcher and al in [7] to cover the case of sufficiently small domains in Riemannian manifolds with nonnegative curvature. A complete generalization to manifolds with positive or negative curvature, including existence and uniqueness results (under some convexity conditions in positive curvature), has been given by one of the authors in [22]. In the case $p = 2$, computation of the Riemannian barycenter has been performed by Le in [13] using a gradient descent algorithm.

In this paper, we consider the general case $p \geq 1$ in a Riemannian setting. Under the above mentioned condition of Asfari [1] ensuring existence and uniqueness of the p-mean, we provide a stochastic gradient descent algorithm that converges almost surely to e_p. This algorithm is easier to implement than the deterministic gradient descent algorithm since it does not require computing the gradient of the functional H_p to minimize. More precisely, we construct a time inhomogeneous Markov chain $(X_k)_{k\geq 0}$ as follows: at each step $k \geq 0$, we draw a random point P_{k+1} with distribution μ and we move the current point X_k to X_{k+1} along the geodesic from X_k to X_{k+1} by a distance depending on p, X_k and on a deteministic parameter t_{k+1}. In Theorem 2.3 below, we state that under a suitable condition on the sequence $(t_k)_{k\geq 1}$, the Markov chain $(X_k)_{k\geq 0}$ converges almost surely and in L^2 to the p-mean e_p. Our proof relies on the martingale convergence theorem and the main point consists in determining and estimating all the geometric quantities. For related convergence results on recursive stochastic algorithms, see [14] Theorem 1 or [3].

We then study the speed of convergence to the p-mean and its fluctuations: in Theorem 2.6 we prove that the suitably renormalized inhomogeneous Markov chain $(X_k)_{k\geq 0}$ converges in law to an inhomogeneous diffusion process in the Skohorod space. This is an invariance principle type result, see e.g. [9], [15], [4], [6] for related works. Interestingly, the limiting process depends in a crucial way on the sequence $(t_k)_{k\geq 1}$ of the algorithm. The main point is to compute the generator of the rescaled Markov chain and to obtain the characteristics of the limiting process from the curvature conditions and estimates on Jacobi fields.

The paper is organized as follows. Section 2 is devoted to a detailed presentation of the stochastic gradient descent algorithm $(X_k)_{k\geq 0}$ and its properties: almost sure convergence is stated in Theorem 2.3 and the invariance principle in Theorem 2.6. Proofs are gathered in Section 3.

2. Results

2.1. p-means in regular geodesic balls. Let M be a Riemannian manifold with pinched sectional curvatures. Let $\alpha, \beta > 0$ such that α^2 is a positive upper bound for sectional curvatures on M, and $-\beta^2$ is a negative lower bound for sectional curvatures on M. Denote by ρ the Riemannian distance on M.

In M consider a geodesic ball $B(a,r)$ with $a \in M$. Let μ be a probability measure with support included in a compact convex subset K_μ of $B(a,r)$. Fix $p \in [1,\infty)$. We will always make the following assumptions on (r,p,μ):

Assumption 2.1. The support of μ is not reduced to one point. Either $p > 1$ or the support of μ is not contained in a line, and the radius r satisfies equation (1.2).

Note that $B(a,r)$ is convex if $r < \frac{1}{2}\min\left\{\mathrm{inj}(M), \frac{\pi}{\alpha}\right\}$. Under assumption 2.1, it has been proved in [1] (Theorem 2.1) that the functional H_p defined by Equation (1.1) has a unique minimizer e_p in M, the p-mean of μ, and moreover $e_p \in B(a,r)$. If $p = 1$, e_1 is the median of μ. It is easily checked that if $p \in [1,2)$, then H_p is strictly convex on $B(a,r)$. On the other hand, if $p \geq 2$ then H_p is of class C^2 on $B(a,r)$ but not necessarily convex as mentioned in the introduction.

Proposition 2.2. *Let K be a convex subset of $B(a,r)$ containing the support of μ. Then there exists $C_{p,\mu,K} > 0$ such that for all $x \in K$,*

$$H_p(x) - H_p(e_p) \geq \frac{C_{p,\mu,K}}{2}\rho(x,e_p)^2. \tag{2.1}$$

Moreover if $p \geq 2$ then we can choose $C_{p,\mu,K}$ so that for all $x \in K$,

$$\|\operatorname{grad}_x H_p\|^2 \geq C_{p,\mu,K}\left(H_p(x) - H_p(e_p)\right). \tag{2.2}$$

In the sequel, we fix

$$K = \bar{B}(a, r-\varepsilon) \quad \text{with} \quad \varepsilon = \frac{\rho(K_\mu, B(a,r)^c)}{2}. \tag{2.3}$$

We now state our main result: we define a stochastic gradient algorithm $(X_k)_{k\geq 0}$ to approximate the p-mean e_p and prove its convergence.

Theorem 2.3. *Let $(P_k)_{k\geq 1}$ be a sequence of independent $B(a,r)$-valued random variables, with law μ. Let $(t_k)_{k\geq 1}$ be a sequence of positive numbers satisfying*

$$\forall k \geq 1, \quad t_k \leq \min\left(\frac{1}{C_{p,\mu,K}}, \frac{\rho(K_\mu, B(a,r)^c)}{2p(2r)^{p-1}}\right), \tag{2.4}$$

$$\sum_{k=1}^{\infty} t_k = +\infty \quad \textit{and} \quad \sum_{k=1}^{\infty} t_k^2 < \infty. \tag{2.5}$$

Letting $x_0 \in K$, define inductively the random walk $(X_k)_{k\geq 0}$ by

$$X_0 = x_0 \quad \textit{and for } k \geq 0 \quad X_{k+1} = \exp_{X_k}\left(-t_{k+1}\operatorname{grad}_{X_k} F_p(\cdot, P_{k+1})\right) \tag{2.6}$$

where $F_p(x,y) = \rho^p(x,y)$, with the convention $\operatorname{grad}_x F_p(\cdot, x) = 0$.

The random walk $(X_k)_{k\geq 1}$ converges in L^2 and almost surely to e_p.

In the following example, we focus on the case $M = \mathbb{R}^d$ and $p = 2$ where drastic simplifications occur.

Example 2.4. In the case when $M = \mathbb{R}^d$ and μ is a compactly supported probability measure on $\mathbb{R}^d$, the stochastic gradient algorithm (2.6) simplifies into

$$X_0 = x_0 \quad \text{and for } k \geq 0 \quad X_{k+1} = X_k - t_{k+1} \operatorname{grad}_{X_k} F_p(\cdot, P_{k+1}).$$

If furthermore $p = 2$, clearly $e_2 = \mathbb{E}[P_1]$ and $\operatorname{grad}_x F_p(\cdot, y) = 2(x - y)$, so that the linear relation

$$X_{k+1} = (1 - 2t_{k+1})X_k + 2t_{k+1}P_{k+1}, \quad k \geq 0$$

holds true and an easy induction proves that

$$X_k = x_0 \prod_{j=0}^{k-1}(1 - 2t_{k-j}) + 2\sum_{j=0}^{k-1} P_{k-j}t_{k-j} \prod_{\ell=0}^{j-1}(1 - 2t_{k-\ell}), \quad k \geq 1. \tag{2.7}$$

Now, taking $t_k = \dfrac{1}{2k}$, we have

$$\prod_{j=0}^{k-1}(1 - 2t_{k-j}) = 0 \quad \text{and} \quad \prod_{\ell=0}^{j-1}(1 - 2t_{k-\ell}) = \frac{k-j}{k}$$

so that

$$X_k = \sum_{j=0}^{k-1} P_{k-j}\frac{1}{k} = \frac{1}{k}\sum_{j=1}^{k} P_j.$$

The stochastic gradient algorithm estimating the mean e_2 of μ is given by the empirical mean of a growing sample of independent random variables with distribution μ. In this simple case, the result of Theorem 2.3 is nothing but the strong law of large numbers. Moreover, fluctuations around the mean are given by the central limit theorem and Donsker's theorem.

2.2. **Fluctuations of the stochastic gradient algorithm.** The notations are the same as in the beginning of section 2.1. We still make assumption 2.1. Let us define K and ε as in (2.3) and let

$$\delta_1 = \min\left(\frac{1}{C_{p,\mu,K}}, \frac{\rho(K_\mu, B(a,r)^c)}{2p(2r)^{p-1}}\right). \tag{2.8}$$

We consider the time inhomogeneous M-valued Markov chain (2.6) in the particular case when

$$t_k = \min\left(\frac{\delta}{k}, \delta_1\right), \quad k \geq 1 \tag{2.9}$$

for some $\delta > 0$. The particular sequence $(t_k)_{k\geq1}$ defined by (2.9) satisfies (2.4) and (2.5), so Theorem 2.3 holds true and the stochastic gradient algorithm $(X_k)_{k\geq0}$ converges a.s. and in L^2 to the p-mean e_p.

In order to study the fluctuations around the p-mean e_p, we define for $n \geq 1$ the rescaled $T_{e_p}M$-valued Markov chain $(Y_k^n)_{k\geq0}$ by

$$Y_k^n = \frac{k}{\sqrt{n}} \exp_{e_p}^{-1} X_k. \tag{2.10}$$

We will prove convergence of the sequence of process $(Y_{[nt]}^n)_{t\geq0}$ to a non-homogeneous diffusion process. The limit process is defined in the following proposition:

Proposition 2.5. *Assume that H_p is C^2 in a neighborhood of e_p, and that $\delta > C^{-1}_{p,\mu,K}$. Define*

$$\Gamma = \mathbb{E}\left[\operatorname{grad}_{e_p} F_p(\cdot, P_1) \otimes \operatorname{grad}_{e_p} F_p(\cdot, P_1)\right]$$

and $G_\delta(t)$ the generator

$$G_\delta(t)f(y) := \langle d_y f, t^{-1}(y - \delta\nabla dH_p(y,\cdot)^\sharp)\rangle + \frac{\delta^2}{2}\mathrm{Hess}_y f\,(\Gamma) \tag{2.11}$$

where $\nabla dH_p(y,\cdot)^\sharp$ denotes the dual vector of the linear form $\nabla dH_p(y,\cdot)$.

There exists a unique inhomogeneous diffusion process $(y_\delta(t))_{t>0}$ on $T_{e_p}M$ with generator $G_\delta(t)$ and converging in probability to 0 as $t \to 0^+$.

The process y_δ is continuous, converges a.s. to 0 as $t \to 0^+$ and has the following integral representation:

$$y_\delta(t) = \sum_{i=1}^d t^{1-\delta\lambda_i}\int_0^t s^{\delta\lambda_i - 1}\langle \delta\sigma\, dB_s, e_i\rangle e_i, \quad t \geq 0, \tag{2.12}$$

where B_t is a standard Brownian motion on $T_{e_p}M$, $\sigma \in \mathrm{End}(T_{e_p}M)$ satisfies $\sigma\sigma^* = \Gamma$, $(e_i)_{1\leq i\leq d}$ is an orthonormal basis diagonalizing the symmetric bilinear form $\nabla dH_p(e_p)$ and $(\lambda_i)_{1\leq i\leq d}$ are the associated eigenvalues.

Note that the integral representation (2.12) implies that y_δ is the centered Gaussian process with covariance

$$\mathbb{E}\left[y_\delta^i(t_1)y_\delta^j(t_2)\right] = \frac{\delta^2\Gamma(e_i^*\otimes e_j^*)}{\delta(\lambda_i+\lambda_j)-1}t_1^{1-\delta\lambda_i}t_2^{1-\delta\lambda_j}(t_1\wedge t_2)^{\delta(\lambda_i+\lambda_j)-1}, \tag{2.13}$$

where $y_\delta^i(t) = \langle y_\delta(t), e_i\rangle$, $1 \leq i, j \leq d$ and $t_1, t_2 \geq 0$.

Our main result on the fluctuations of the stochastic gradient algorithm is the following:

Theorem 2.6. *Assume that either e_p does not belong to the support of μ or $p \geq 2$. Assume furthermore that $\delta > C^{-1}_{p,\mu,K}$. The sequence of processes $\left(Y^n_{[nt]}\right)_{t\geq 0}$ weakly converges in $\mathbb{D}((0,\infty), T_{e_p}M)$ to y_δ.*

Remark 2.7. The assumption on e_p implies that H_p is of class C^2 in a neighbourhood of e_p. For most of the applications μ is equidistributed on a finite set of data which can be considered as randomly distributed. In this situation, when $p > 1$ then almost surely e_p does not belong to the support of μ. For $p = 1$ one has to be more careful since with positive probability e_1 belongs to the support of μ.

Remark 2.8. From section 2.1 we know that, when $p \in (1, 2]$, the constant

$$C_{p,\mu,K} = p(2r)^{p-2}\left(\min\left(p-1, 2\alpha r\cot\left(2\alpha r\right)\right)\right)$$

is explicit. The constraint $\delta > C^{-1}_{p,\mu,K}$ can easily be checked in this case.

Remark 2.9. In the case $M = \mathbb{R}^d$, $Y_k^n = \frac{k}{\sqrt{n}}(X_k - e_p)$ and the tangent space $T_{e_p}M$ is identified to $\mathbb{R}^d$. Theorem 2.6 holds and, in particular, when $t = 1$, we obtain a central limit Theorem: $\sqrt{n}(X_n - e_p)$ converges as $n \to \infty$ to a centered Gaussian d-variate distribution (with covariance structure given by (2.13) with $t_1 = t_2 = 1$). This is a central limit theorem: the fluctuations of the stochastic gradient algorithm are of scale $n^{-1/2}$ and asymptotically Gaussian.

3. Proofs

For simplicity, let us write shortly $e = e_p$ in the proofs.

3.1. **Proof of Proposition 2.2.**

For $p = 1$ this is a direct consequence of [22] Theorem 3.7.

Next we consider the case $p \in (1, 2)$.

Let $K \subset B(a, r)$ be a compact convex set containing the support of μ. Let $x \in K \backslash \{e\}$, $t = \rho(e, x)$, $u \in T_eM$ the unit vector such that $\exp_e(\rho(e, x)u) = x$, and γ_u the geodesic with initial speed u : $\dot\gamma_u(0) = u$. For $y \in K$, letting $h_y(s) = \rho(\gamma_u(s), y)^p$, $s \in [0, t]$, we have since $p > 1$

$$h_y(t) = h_y(0) + th'_y(0) + \int_0^t (t - s)h''_y(s)\, ds$$

with the convention $h''_y(s) = 0$ when $\gamma_u(s) = y$. Indeed, if $y \notin \gamma([0, t])$ then h_y is smooth, and if $y \in \gamma([0, t])$, say $y = \gamma(s_0)$ then $h_y(s) = |s - s_0|^p$ and the formula can easily be checked.

By standard calculation,

$$\begin{aligned} &h''_y(s) \\ &\geq p\rho(\gamma_u(s), y)^{p-2} \\ &\quad\times \Big((p-1)\|\dot\gamma_u(s)^{T(y)}\|^2 + \|\dot\gamma_u(s)^{N(y)}\|^2 \alpha\rho(\gamma_u(s), y) \cot(\alpha\rho(\gamma_u(s), y))\Big) \end{aligned} \tag{3.1}$$

with $\dot\gamma_u(s)^{T(y)}$ (resp. $\dot\gamma_u(s)^{N(y)}$) the tangential (resp. the normal) part of $\dot\gamma_u(s)$ with respect to $n(\gamma_u(s), y) = \dfrac{1}{\rho(\gamma_u(s), y)} \exp^{-1}_{\gamma_u(s)}(y)$:

$$\dot\gamma_u(s)^{T(y)} = \langle \dot\gamma_u(s), n(\gamma_u(s), y)\rangle n(\gamma_u(s), y), \quad \dot\gamma_u(s)^{N(y)} = \dot\gamma_u(s) - \dot\gamma_u(s)^{T(y)}.$$

From this we get

$$h''_y(s) \geq p\rho(\gamma_u(s), y)^{p-2} \left(\min\left(p - 1, 2\alpha r \cot(2\alpha r)\right)\right). \tag{3.2}$$

Now

$$\begin{aligned} &H_p(\gamma_u(t')) \\ &= \int_{B(a,r)} h_y(\gamma_u(t'))\, \mu(dy) \\ &= \int_{B(a,r)} h_y(0)\, \mu(dy) + t' \int_{B(a,r)} h'_y(0)\, \mu(dy) + \int_0^{t'} (t' - s) \left(\int_{B(a,r)} h_y(s)''\, \mu(dy) \right) ds \end{aligned}$$

and $H_p(\gamma_u(t'))$ attains its minimum at $t' = 0$, so $\displaystyle\int_{B(a,r)} h'_y(0)\, \mu(dy) = 0$ and we have

$$H_p(x) = H_p(\gamma_u(t)) = H_p(e) + \int_0^t (t - s) \left(\int_{B(a,r)} h_y(s)''\, \mu(dy) \right) ds.$$

Using Equation (3.2) we get

(3.3)
$$H_p(x) \geq H_p(e) + \int_0^t \left((t-s) \int_{B(a,r)} p\rho(\gamma_u(s), y)^{p-2} \left(\min\left(p-1, 2\alpha r \cot(2\alpha r)\right)\right) \mu(dy) \right) ds.$$

Since $p \leq 2$ we have $\rho(\gamma_u(s), y)^{p-2} \geq (2r)^{p-2}$ and

$$H_p(x) \geq H_p(e) + \frac{t^2}{2} p(2r)^{p-2} \left(\min\left(p-1, 2\alpha r \cot(2\alpha r)\right)\right). \tag{3.4}$$

So letting

$$C_{p,\mu,K} = p(2r)^{p-2} \left(\min\left(p-1, 2\alpha r \cot(2\alpha r)\right)\right)$$

we obtain

$$H_p(x) \geq H_p(e) + \frac{C_{p,\mu,K}\,\rho(e,x)^2}{2}. \tag{3.5}$$

To finish let us consider the case $p \geq 2$.

In the proof of [1] Theorem 2.1, it is shown that e is the only zero of the maps $x \mapsto \operatorname{grad}_x H_p$ and $x \mapsto H_p(x) - H_p(e)$, and that $\nabla dH_p(e)$ is strictly positive. This implies that (2.1) and (2.2) hold on some neighbourhood $B(e, \varepsilon)$ of e. By compactness and the fact that $H_p - H_p(e)$ and $\operatorname{grad} H_p$ do not vanish on $K \backslash B(e, \varepsilon)$ and $H_p - H_p(e)$ is bounded, possibly modifying the constant $C_{p,\mu,K}$, (2.1) and (2.2) also holds on $K \backslash B(e, \varepsilon)$.

□

3.2. **Proof of Theorem 2.3.**
Note that, for $x \neq y$,

$$\operatorname{grad}_x F(\cdot, y) = p\rho^{p-1}(x,y) \frac{-\exp_x^{-1}(y)}{\rho(x,y)} = -p\rho^{p-1}(x,y) n(x,y),$$

whith $n(x,y) := \dfrac{\exp_x^{-1}(y)}{\rho(x,y)}$ a unit vector. So, with the condition (2.4) on t_k, the random walk $(X_k)_{k \geq 0}$ cannot exit K: if $X_k \in K$ then there are two possibilities for X_{k+1}:

- either X_{k+1} is in the geodesic between X_k and P_{k+1} and belongs to K by convexity of K;
- or X_{k+1} is after P_{k+1}, but since

$$\begin{aligned} \| t_{k+1} \operatorname{grad}_{X_k} F_p(\cdot, P_{k+1}) \| &= t_{k+1} p\rho^{p-1}(X_k, P_{k+1}) \\ &\leq \frac{\rho(K_\mu, B(a,r)^c)}{2p(2r)^{p-1}} p\rho^{p-1}(X_k, P_{k+1}) \\ &\leq \frac{\rho(K_\mu, B(a,r)^c)}{2}, \end{aligned}$$

  we have in this case

$$\rho(P_{k+1}, X_{k+1}) \leq \frac{\rho(K_\mu, B(a,r)^c)}{2}$$

  which implies that $X_{k+1} \in K$.

First consider the case $p \in [1,2)$.

For $k \geq 0$ let

$$t \mapsto E(t) := \frac{1}{2}\rho^2\left(e, \gamma(t)\right),$$

$\gamma(t)_{t\in[0,t_{k+1}]}$ the geodesic satisfying $\dot\gamma(0) = -\operatorname{grad}_{X_k} F_p(\cdot, P_{k+1})$. We have for all $t \in [0, t_{k+1}]$

$$E''(t) \leq C(\beta, r, p) := p^2(2r)^{2p-1}\beta \operatorname{cotanh}(2\beta r) \tag{3.6}$$

(see e.g. [22]). By Taylor formula,

$$\begin{aligned}
&\rho(X_{k+1}, e)^2\\
&= 2E(t_{k+1})\\
&= 2E(0) + 2t_{k+1}E'(0) + t_{k+1}^2 E''(t) \quad \text{for some } t \in [0, t_{k+1}]\\
&\leq \rho(X_k, e)^2 + 2t_{k+1}\langle \operatorname{grad}_{X_k} F_p(\cdot, P_{k+1}), \exp_{X_k}^{-1}(e)\rangle + t_{k+1}^2 C(\beta, r, p).
\end{aligned}$$

Now from the convexity of $x \mapsto F_p(x,y)$ we have for all $x, y \in B(a,r)$

$$F_p(e,y) - F_p(x,y) \geq \left\langle \operatorname{grad}_x F_p(\cdot, y), \exp_x^{-1}(e)\right\rangle. \tag{3.7}$$

This applied with $x = X_k$, $y = P_{k+1}$ yields

$$\rho(X_{k+1}, e)^2 \leq \rho(X_k, e)^2 - 2t_{k+1}\left(F_p(X_k, P_{k+1}) - F_p(e, P_{k+1})\right) + C(\beta, r, p)t_{k+1}^2. \tag{3.8}$$

Letting for $k \geq 0$ $\mathscr{F}_k = \sigma(X_\ell, 0 \leq \ell \leq k)$, we get

$$\begin{aligned}
&\mathbb{E}\left[\rho(X_{k+1}, e)^2 | \mathscr{F}_k\right]\\
&\leq \rho(X_k, e)^2 - 2t_{k+1}\int_{B(a,r)} \left(F_p(X_k, y) - F_p(e, y)\right)\mu(dy) + C(\beta, r, p)t_{k+1}^2\\
&= \rho(X_k, e)^2 - 2t_{k+1}\left(H_p(X_k) - H_p(e)\right) + C(\beta, r, p)t_{k+1}^2\\
&\leq \rho(X_k, e)^2 + C(\beta, r, p)t_{k+1}^2
\end{aligned}$$

so that the process $(Y_k)_{k\geq 0}$ defined by

$$Y_0 = \rho(X_0, e)^2 \quad \text{and for } k \geq 1 \quad Y_k = \rho(X_k, e)^2 - C(\beta, r, p)\sum_{j=1}^k t_j^2 \tag{3.9}$$

is a bounded supermartingale. So it converges in L^1 and almost surely. Consequently $\rho(X_k, e)^2$ also converges in L^1 and almost surely.

Let

$$a = \lim_{k\to\infty} \mathbb{E}\left[\rho(X_k, e)^2\right]. \tag{3.10}$$

We want to prove that $a = 0$. We already proved that

$$\mathbb{E}\left[\rho(X_{k+1}, e)^2 | \mathscr{F}_k\right] \leq \rho(X_k, e)^2 - 2t_{k+1}\left(H_p(X_k) - H_p(e)\right) + C(\beta, r, p)t_{k+1}^2. \tag{3.11}$$

Taking the expectation and using Proposition 2.2, we obtain

$$\mathbb{E}\left[\rho(X_{k+1}, e)^2\right] \leq \mathbb{E}\left[\rho(X_k, e)^2\right] - t_{k+1}C_{p,\mu,K}\mathbb{E}\left[\rho(X_k, e)^2\right] + C(\beta, r, p)t_{k+1}^2. \tag{3.12}$$

An easy induction proves that for $\ell \geq 1$,

$$\mathbb{E}\left[\rho(X_{k+\ell}, e)^2\right] \leq \prod_{j=1}^{\ell}(1 - C_{p,\mu,K}t_{k+j})\mathbb{E}\left[\rho(X_k, e)^2\right] + C(\beta, r, p)\sum_{j=1}^{\ell} t_{k+j}^2. \tag{3.13}$$

Letting $\ell \to \infty$ and using the fact that $\sum_{j=1}^{\infty} t_{k+j} = \infty$ which implies

$$\prod_{j=1}^{\infty}(1 - C_{p,\mu,K} t_{k+j}) = 0,$$

we get

$$a \le C(\beta, r, p) \sum_{j=1}^{\infty} t_{k+j}^2. \tag{3.14}$$

Finally using $\sum_{j=1}^{\infty} t_j^2 < \infty$ we obtain that $\lim_{k\to\infty} \sum_{j=1}^{\infty} t_{k+j}^2 = 0$, so $a = 0$. This proves L^2 and almost sure convergence.

Next assume that $p \ge 2$.

For $k \ge 0$ let

$$t \mapsto E_p(t) := H_p(\gamma(t)),$$

$\gamma(t)_{t\in[0,t_{k+1}]}$ the geodesic satisfying $\dot\gamma(0) = -\operatorname{grad}_{X_k} F_p(\cdot, P_{k+1})$. With a calculation similar to (3.6) we get for all $t \in [0, t_{k+1}]$

$$E_p''(t) \le 2C(\beta, r, p) := p^3 (2r)^{3p-4} \left(2r\beta \operatorname{cotanh}(2\beta r) + p - 2\right). \tag{3.15}$$

(see e.g. [22]). By Taylor formula,

$$\begin{aligned}
H_p(X_{k+1}) &= E_p(t_{k+1}) \\
&= E_p(0) + t_{k+1} E_p'(0) + \frac{t_{k+1}^2}{2} E_p''(t) \quad \text{for some } t \in [0, t_{k+1}] \\
&\le H_p(X_k) + t_{k+1} \langle d_{X_k} H_p, \operatorname{grad}_{X_k} F_p(\cdot, P_{k+1})\rangle + t_{k+1}^2 C(\beta, r, p).
\end{aligned}$$

We get

$$\begin{aligned}
&\mathbb{E}\left[H_p(X_{k+1}) | \mathscr{F}_k\right] \\
&\le H_p(X_k) - t_{k+1} \left\langle d_{X_k} H_p, \int_{B(a,r)} \operatorname{grad}_{X_k} F_p(\cdot, y)\mu(dy) \right\rangle + C(\beta, r, p) t_{k+1}^2 \\
&= H_p(X_k) - t_{k+1} \left\langle d_{X_k} H_p, \operatorname{grad}_{X_k} H_p(\cdot) \right\rangle + C(\beta, r, p) t_{k+1}^2 \\
&= H_p(X_k) - t_{k+1} \left\| \operatorname{grad}_{X_k} H_p(\cdot) \right\|^2 + C(\beta, r, p) t_{k+1}^2 \\
&\le H_p(X_k) - C_{p,\mu,K} t_{k+1} \left(H_p(X_k) - H_p(e)\right) + C(\beta, r, p) t_{k+1}^2
\end{aligned}$$

(by Proposition 1.1) so that the process $(Y_k)_{k\ge0}$ defined by

$$Y_0 = H_p(X_0) - H_p(e) \quad \text{and for } k \ge 1 \quad Y_k = H_p(X_k) - H_p(e) - C(\beta, r, p) \sum_{j=1}^{k} t_j^2 \tag{3.16}$$

is a bounded supermartingale. Now the argument is exactly the same as in the first part to prove that $H_p(X_k) - H_p(e)$ also converges in L^1 and almost surely to 0.

Finally (2.1) proves that $\rho(X_k, e)^2$ converges in L^1 and almost surely to 0. $\square$

3.3. **Proof of Proposition 2.5.** Fix $\varepsilon > 0$. Any diffusion process on $[\varepsilon, \infty)$ with generator $G_\delta(t)$ is solution of a sde of the type

$$dy_t = \frac{1}{t} L_\delta(y_t)\, dt + \delta\sigma\, dB_t \tag{3.17}$$

where $L_\delta(y) = y - \delta\nabla dH_p(y, \cdot)^\sharp$ and B_t and σ are as in Proposition 2.5. This sde can be solved explicitely on $[\varepsilon, \infty)$. The symmetric endomorphism $y \mapsto \nabla dH_p(y, \cdot)^\sharp$ is diagonalisable in the orthonormal basis $(e_i)_{1\le i\le d}$ with eigenvalues $(\lambda_i)_{1\le i\le d}$. The endomorphism $L_\delta = \mathrm{id} - \delta\nabla dH_p(e)(\mathrm{id}, \cdot)^\sharp$ is also diagonalisable in this basis with eigenvalues $(1 - \delta\lambda_i)_{1\le i\le d}$. The solution $y_t = \sum_{i=1}^d y_t^i e_i$ of (3.17) started at $y_\varepsilon = \sum_{i=1}^d y_\varepsilon^i e_i$ is given by

$$y_t = \sum_{i=1}^d \left(y_\varepsilon^i \varepsilon^{\delta\lambda_i - 1} + \int_\varepsilon^t s^{\delta\lambda_i - 1} \langle \delta\sigma\, dB_s, e_i\rangle \right) t^{1-\delta\lambda_i} e_i, \quad t \ge \varepsilon. \tag{3.18}$$

Now by definition of $C_{p,\mu,K}$ we clearly have

$$C_{p,\mu,K} \le \min_{1\le i\le d} \lambda_i. \tag{3.19}$$

So the condition $\delta C_{p,\mu,K} > 1$ implies that for all i, $\delta\lambda_i - 1 > 0$, and as $\varepsilon \to 0$,

$$\int_\varepsilon^t s^{\delta\lambda_i - 1}\langle \delta\sigma\, dB_s, e_i\rangle \to \int_0^t s^{\delta\lambda_i - 1}\langle \delta\sigma\, dB_s, e_i\rangle \quad \text{in probability.} \tag{3.20}$$

Assume that a continuous solution y_t converging in probability to 0 as $t \to 0^+$ exists. Since $y_\varepsilon^i \varepsilon^{\delta\lambda_i - 1} \to 0$ in probability as $\varepsilon \to 0$, we necessarily have using (3.20)

$$y_t = \sum_{i=1}^d t^{1-\delta\lambda_i} \int_0^t s^{\delta\lambda_i - 1}\langle \delta\sigma\, dB_s, e_i\rangle e_i, \quad t \ge 0. \tag{3.21}$$

Note y_δ^i is Gaussian with variance $\dfrac{t\delta^2\Gamma(e_i^* \otimes e_i^*)}{2\delta\lambda_i - 1}$, so it converges in L^2 to 0 as $t \to 0$. Conversely, it is easy to check that equation (3.21) defines a solution to (3.17).

To prove the a.s. convergence to 0 we use the representation

$$\int_0^t s^{\delta\lambda_i - 1}\langle \delta\sigma\, dB_s, e_i\rangle = B_{\varphi_i(t)}^i$$

where B_s^i is a Brownian motion and $\varphi_i(t) = \dfrac{\delta^2\Gamma(e_i^* \otimes e_i^*)}{2\delta\lambda_i - 1} t^{2\delta\lambda_i - 1}$. Then by the law of iterated logarithm

$$\limsup_{t\downarrow 0} t^{1-\delta\lambda_i} B_{\varphi_i(t)}^i \le \limsup_{t\downarrow 0} t^{1-\delta\lambda_i} \sqrt{2\varphi_i(t) \ln\ln\left(\varphi_i^{-1}(t)\right)}$$

But for t small we have

$$\sqrt{2\varphi_i(t)\ln\ln\left(\varphi_i^{-1}(t)\right)} \le t^{\delta\lambda_i - 3/4}$$

so

$$\limsup_{t\downarrow 0} t^{1-\delta\lambda_i} B_{\varphi_i(t)}^i \le \lim_{t\downarrow 0} t^{1/4} = 0.$$

This proves a.s. convergence to 0. Continuity is easily checked using the integral representation (3.21). $\square$

3.4. **Proof of Theorem 2.6.** Consider the time homogeneous Markov chain $(Z_k^n)_{k\ge 0}$ with state space $[0,\infty)\times T_eM$ defined by

$$Z_k^n = \left(\frac{k}{n}, Y_k^n\right). \tag{3.22}$$

The first component has a deterministic evolution and will be denoted by t_k^n; it satisfies

$$t_{k+1}^n = t_k^n + \frac{1}{n}, \quad k\ge 0. \tag{3.23}$$

Let k_0 be such that

$$\frac{\delta}{k_0} < \delta_1. \tag{3.24}$$

Using equations (2.6), (2.10) and (2.9), we have for $k\ge k_0$,

$$Y_{k+1}^n = \frac{nt_k^n+1}{\sqrt{n}}\exp_e^{-1}\left(\exp_{\exp_e\frac{1}{\sqrt{nt_k^n}}Y_k^n}\left(-\frac{\delta}{nt_k^n+1}\operatorname{grad}_{\frac{1}{\sqrt{nt_k^n}}Y_k^n}F_p(\cdot,P_{k+1})\right)\right). \tag{3.25}$$

Consider the transition kernel $P^n(z,dz')$ on $(0,\infty)\times T_eM$ defined for $z=(t,y)$ by

$$P^n(z,A) = \mathbb{P}\left[\left(t+\frac{1}{n},\frac{nt+1}{\sqrt{n}}\exp_e^{-1}\left(\exp_{\exp_e\frac{1}{\sqrt{nt}}y}\left(-\frac{\delta}{nt+1}\operatorname{grad}_{\exp_e\frac{1}{\sqrt{nt}}y}F_p(\cdot,P_1)\right)\right)\right)\in A\right] \tag{3.26}$$

where $A\in\mathscr{B}((0,\infty)\times T_eM)$. Clearly this transition kernel drives the evolution of the Markov chain $(Z_k^n)_{k\ge k_0}$.

For the sake of clarity, we divide the proof of Theorem 2.6 into four lemmas.

Lemma 3.1. *Assume that either $p\ge 2$ or e does not belong to the support* $\operatorname{supp}(\mu)$ *of μ (note this implies that for all $x\in\operatorname{supp}(\mu)$ the function $F_p(\cdot,x)$ is of class C^2 in a neighbourhood of e). Fix $\delta>0$. Let B be a bounded set in T_eM and let $0<\varepsilon<T$. We have for all C^2 function f on T_eM*

$$\begin{aligned}
&n\left(f\left(\frac{nt+1}{\sqrt{n}}\exp_e^{-1}\left(\exp_{\exp_e\frac{1}{\sqrt{nt}}y}\left(-\frac{\delta}{nt+1}\operatorname{grad}_{\exp_e\frac{1}{\sqrt{nt}}y}F_p(\cdot,x)\right)\right)\right)-f(y)\right)\\
&=\left\langle d_yf,\frac{y}{t}\right\rangle-\sqrt{n}\langle d_yf,\delta\operatorname{grad}_eF_p(\cdot,x)\rangle-\delta\nabla dF_p(\cdot,x)\left(\operatorname{grad}_y f,\frac{y}{t}\right)\\
&+\frac{\delta^2}{2}\operatorname{Hess}_y f\left(\operatorname{grad}_eF_p(\cdot,x)\otimes\operatorname{grad}_eF_p(\cdot,x)\right)+O\left(\frac{1}{\sqrt{n}}\right)
\end{aligned} \tag{3.27}$$

uniformly in $y\in B$, $x\in\operatorname{supp}(\mu)$, $t\in[\varepsilon,T]$.

Proof. Let $x\in\operatorname{supp}(\mu)$, $y\in T_eM$, $u,v\in\mathbb{R}$ sufficiently close to 0, and $q=\exp_e\left(\frac{uy}{t}\right)$. For $s\in[0,1]$ denote by $a\mapsto c(a,s,u,v)$ the geodesic with endpoints $c(0,s,u,v)=e$ and

$$c(1,s,u,v)=\exp_{\exp_e\left(\frac{uy}{t}\right)}\left(-vs\operatorname{grad}_{\exp_e\left(\frac{uy}{t}\right)}F_p(\cdot,x)\right):$$

$$c(a,s,u,v) = \exp_e\left\{a\exp_e^{-1}\left[\exp_{\exp_e\left(\frac{uy}{t}\right)}\left(-sv\,\mathrm{grad}_{\exp_e\left(\frac{uy}{t}\right)}F_p(\cdot,x)\right)\right]\right\}.$$

This is a C^2 function of $(a,s,u,v)\in[0,1]^2\times(-\eta,\eta)^2$, η sufficiently small. It also depends in a C^2 way of x and y. Letting $c(a,s) = c\left(a,s,\frac{1}{\sqrt{n}},\frac{\delta}{nt+1}\right)$, we have

$$\exp_e^{-1}\left(\exp_{\exp_e\frac{1}{\sqrt{nt}}y}\left(-\frac{\delta}{nt+1}\,\mathrm{grad}_{\exp_e\frac{1}{\sqrt{nt}}y}F_p(\cdot,x)\right)\right) = \partial_a c(0,1).$$

So we need a Taylor expansion up to order n^{-1} of $\frac{nt+1}{\sqrt{n}}\partial_a c(0,1)$.

We have $c(a,s,0,1) = \exp_e\left(-as\,\mathrm{grad}_e F_p(\cdot,x)\right)$ and this implies

$$\partial_s^2\partial_a c(0,s,0,1) = 0, \quad \text{so} \quad \partial_s^2\partial_a c(0,s,u,1) = O(u).$$

On the other hand the identities $c(a,s,u,v) = c(a,sv,u,1)$ yields $\partial_s^2\partial_a c(a,s,u,v) = v^2\partial_s^2\partial_a c(a,s,u,1)$, so we obtain

$$\partial_s^2\partial_a c(0,s,u,v) = O(uv^2)$$

and this yields

$$\partial_s^2\partial_a c(0,s) = O(n^{-5/2}),$$

uniformly in s,x,y,t. But since

$$\|\partial_a c(0,1) - \partial_a c(0,0) - \partial_s\partial_a c(0,0)\| \le \frac{1}{2}\sup_{s\in[0,1]}\|\partial_s^2\partial_a c(0,s)\|$$

we only need to estimate $\partial_a c(0,0)$ and $\partial_s\partial_a c(0,0)$.

Denoting by $J(a)$ the Jacobi field $\partial_s c(a,0)$ we have

$$\frac{nt+1}{\sqrt{n}}\partial_a c(0,1) = \frac{nt+1}{\sqrt{n}}\partial_a c(0,0) + \frac{nt+1}{\sqrt{n}}\dot J(0) + O\left(\frac{1}{n^2}\right).$$

On the other hand

$$\frac{nt+1}{\sqrt{n}}\partial_a c(0,0) = \frac{nt+1}{\sqrt{n}}\frac{y}{\sqrt{nt}} = y + \frac{y}{nt}$$

so it remains to estimate $\dot J(0)$.

The Jacobi field $a\mapsto J(a,u,v)$ with endpoints $J(0,u,v) = 0_e$ and

$$J(1,u,v) = -v\,\mathrm{grad}_{\exp_e\left(\frac{uy}{t}\right)}F_p(\cdot,x)$$

satisfies

$$\nabla_a^2 J(a,u,v) = -R(J(a,u,v),\partial_a c(a,0,u,v))\partial_a c(a,0,u,v) = O(u^2 v).$$

This implies that

$$\nabla_a^2 J(a) = O(n^{-2}).$$

Consequently, denoting by $P_{x_1,x_2}: T_{x_1}M\to T_{x_2}M$ the parallel transport along the minimal geodesic from x_1 to x_2 (whenever it is unique) we have

$$P_{c(1,0),e}J(1) = J(0) + \dot J(0) + O(n^{-2}) = \dot J(0) + O(n^{-2}). \tag{3.28}$$

But we also have

$$\begin{aligned}P_{c(1,0,u,v),e}J(1,u,v) &= P_{c(1,0,u,v),e}\left(-v\,\mathrm{grad}_{c(1,0,u,v)}F_p(\cdot,x)\right)\\ &= -v\,\mathrm{grad}_e F_p(\cdot,x) - v\nabla_{\partial_a c(0,0,u,v)}\mathrm{grad}_\cdot F_p(\cdot,x) + O(vu^2)\\ &= -v\,\mathrm{grad}_e F_p(\cdot,x) - v\nabla dF_p(\cdot,x)\left(\frac{uy}{t},\cdot\right)^\sharp + O(vu^2)\end{aligned}$$

where we used $\partial_a c(0,0,u,v) = \frac{uy}{t}$ and for vector fields A, B on TM and a C^2 function f_1 on M

$$\begin{aligned}\langle \nabla_{A_e} \operatorname{grad} f_1, B_e\rangle &= A_e\langle \operatorname{grad} f_1, B_e\rangle - \langle \operatorname{grad} f_1, \nabla_{A_e} B\rangle \\ &= A_e\langle df_1, B_e\rangle - \langle df_1, \nabla_{A_e} B\rangle \\ &= \nabla df_1(A_e, B_e)\end{aligned}$$

which implies

$$\nabla_{A_e} \operatorname{grad} f_1 = \nabla df_1(A_e, \cdot)^\sharp.$$

We obtain

$$P_{c(1,0),e} J(1) = -\frac{\delta}{nt+1} \operatorname{grad}_e F_p(\cdot, x) - \frac{\delta}{\sqrt{n}(nt+1)} \nabla dF_p(\cdot, x)\left(\frac{y}{t}, \cdot\right)^\sharp + O(n^{-2}).$$

Combining with (3.28) this gives

$$\dot{J}(0) = -\frac{\delta}{nt+1} \operatorname{grad}_e F_p(\cdot, x) - \frac{\delta}{nt+1} \nabla dF_p(\cdot, x)\left(\frac{y}{\sqrt{n}t}, \cdot\right)^\sharp + O\left(\frac{1}{n^2}\right).$$

So finally

(3.29)

$$\frac{nt+1}{\sqrt{n}} \partial_a c(0,1) = y + \frac{y}{nt} - \frac{\delta}{\sqrt{n}} \operatorname{grad}_e F_p(\cdot, x) - \delta \nabla dF_p(\cdot, x)\left(\frac{y}{nt}, \cdot\right)^\sharp + O\left(n^{-3/2}\right).$$

To get the final result we are left to make a Taylor expansion of f up to order 2. $\square$

Define the following quantities:

$$b_n(z) = n \int_{\{|z'-z| \le 1\}} (z'-z) P^n(z, dz') \tag{3.30}$$

and

$$a_n(z) = n \int_{\{|z'-z| \le 1\}} (z'-z) \otimes (z'-z) P^n(z, dz'). \tag{3.31}$$

The following property holds:

Lemma 3.2. *Assume that either $p \ge 2$ or e does not belong to the support* $\operatorname{supp}(\mu)$.

(1) *For all $R > 0$ and $\varepsilon > 0$, there exists n_0 such that for all $n \ge n_0$ and $z \in [\varepsilon, T] \times B(0_e, R)$, where $B(0_e, R)$ is the open ball in T_eM centered at the origin with radius R,*

$$\int 1_{\{|z'-z|>1\}} P^n(z, dz') = 0. \tag{3.32}$$

(2) *For all $R > 0$ and $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{z \in [\varepsilon, T] \times B(0_e, R)} |b_n(z) - b(z)| = 0 \tag{3.33}$$

with

$$b(z) = \left(1, \frac{1}{t} L_\delta(y)\right) \quad \text{and} \quad L_\delta(y) = y - \delta \nabla dH(y, \cdot)^\sharp. \tag{3.34}$$

(3) *For all $R > 0$ and $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{z\in[\varepsilon,T]\times B(0_e,R)} |a_n(z) - a(z)| = 0 \tag{3.35}$$

with

$$a(z) = \delta^2 \mathrm{diag}(0,\Gamma) \quad \textit{and} \quad \Gamma = \mathbb{E}\left[\mathrm{grad}_e\, F_p(\cdot, P_1) \otimes \mathrm{grad}_e\, F_p(\cdot, P_1)\right]. \tag{3.36}$$

Proof. (1) We use the notation $z = (t, y)$ and $z' = (t', y')$. We have

$$\begin{aligned}
&\int 1_{\{|z'-z|>1\}}\, P^n(z, dz') \\
&= \int 1_{\{\max(|t'-t|,|y'-y|)>1\}} P^n(z, dz') \\
&= \int 1_{\{\max(\frac{1}{n},|y'-y|)>1\}} P^n(z, dz') \\
&= \mathbb{P}\left[\left|\frac{nt+1}{\sqrt{n}} \exp_e^{-1}\left(\exp_{\exp_e \frac{1}{\sqrt{nt}} y}\left(-\frac{\delta}{nt+1}\, \mathrm{grad}_{\exp_e \frac{1}{\sqrt{nt}} y}\, F_p(\cdot, P_1)\right)\right) - y\right| > 1\right].
\end{aligned}$$

On the other hand, since $F_p(\cdot, x)$ is of class C^2 in a neighbourhood of e, we have by (3.29)

$$\left|\frac{nt+1}{\sqrt{n}} \exp_e^{-1}\left(\exp_{\exp_e \frac{1}{\sqrt{nt}} y}\left(-\frac{\delta}{nt+1}\, \mathrm{grad}_{\exp_e \frac{1}{\sqrt{nt}} y}\, F_p(\cdot, P_1)\right)\right) - y\right| \le \frac{C\delta}{\sqrt{n}\varepsilon} \tag{3.37}$$

for some constant $C > 0$.

(2) Equation (3.32) implies that for $n \ge n_0$

$$\begin{aligned}
&b_n(z) \\
&= n \int (z' - z)\, P^n(z, dz') \\
&= n\left(\frac{1}{n}, \mathbb{E}\left[\frac{nt+1}{\sqrt{n}} \exp_e^{-1}\left(\exp_{\exp_e \frac{y}{\sqrt{nt}}}\left(-\frac{\delta}{nt+1}\, \mathrm{grad}_{\exp_e \frac{y}{\sqrt{nt}}}\, F_p(\cdot, P_1)\right)\right)\right] - y\right).
\end{aligned}$$

We have by lemma 3.1

$$\begin{aligned}
&n\left(\frac{nt+1}{\sqrt{n}} \exp_e^{-1}\left(\exp_{\exp_e \frac{1}{\sqrt{nt}} y}\left(-\frac{\delta}{nt+1}\, \mathrm{grad}_{\exp_e \frac{1}{\sqrt{nt}} y}\, F_p(\cdot, P_1)\right)\right) - y\right) \\
&= \frac{1}{t} y - \delta\sqrt{n}\, \mathrm{grad}_e\, F_p(\cdot, P_1) - \delta \nabla dF_p(\cdot, P_1)\left(\frac{1}{t} y, \cdot\right)^\sharp + O\left(\frac{1}{n^{1/2}}\right)
\end{aligned}$$

a.s. uniformly in n, and since

$$\mathbb{E}\left[\delta\sqrt{n}\, \mathrm{grad}_e\, F_p(\cdot, P_1)\right] = 0,$$

this implies that

$$n\left(\mathbb{E}\left[\frac{nt+1}{\sqrt{n}} \exp_e^{-1}\left(\exp_{\exp_e \frac{1}{\sqrt{nt}} y}\left(-\frac{\delta}{nt+1}\, \mathrm{grad}_{\exp_e \frac{1}{\sqrt{nt}} y}\, F_p(\cdot, P_1)\right)\right)\right] - y\right)$$

converges to

$$\frac{1}{t} y - \mathbb{E}\left[\delta \nabla dF_p(\cdot, P_1)\left(\frac{1}{t} y, \cdot\right)^\sharp\right] = \frac{1}{t} y - \delta \nabla dH_p\left(\frac{1}{t} y, \cdot\right)^\sharp. \tag{3.38}$$

Moreover the convergence is uniform in $z \in [\varepsilon, T] \times B(0_e, R)$, so this yields (3.33).

(3) In the same way, using lemma 3.1,

$$
\begin{aligned}
&n \int (y'-y)\otimes (y'-y)\, P^n(z,dz') \\
&= \frac{1}{n}\mathbb{E}\left[\left(-\sqrt{n}\delta \operatorname{grad}_e F_p(\cdot,P_1)\right)\otimes\left(-\sqrt{n}\delta \operatorname{grad}_e F_p(\cdot,P_1)\right)\right] + o(1) \\
&= \delta^2 \mathbb{E}\left[\operatorname{grad}_e F_p(\cdot,P_1)\otimes \operatorname{grad}_e F_p(\cdot,P_1)\right] + o(1)
\end{aligned}
$$

uniformly in $z \in [\varepsilon, T]\times B(0_e,R)$, so this yields (3.35).

□

Lemma 3.3. *Suppose that* $t_n = \dfrac{\delta}{n}$ *for some* $\delta > 0$. *For all* $\delta > C_{p,\mu,K}^{-1}$,

$$
\sup_{n\ge 1} n\mathbb{E}\left[\rho^2(e,X_n)\right] < \infty. \tag{3.39}
$$

Proof. First consider the case $p \in [1,2)$.

We know by (3.12) that there exists some constant $C(\beta,r,p)$ such that

$$
\mathbb{E}\left[\rho^2(e,X_{k+1})\right] \le \mathbb{E}\left[\rho^2(e,X_k)\right]\exp\left(-C_{p,\mu,K}t_{k+1}\right) + C(\beta,r,p)t_{k+1}^2. \tag{3.40}
$$

From this (3.39) is a consequence of Lemma 0.0.1 (case $\alpha > 1$) in [16]. We give the proof for completeness. We deduce easily by induction that for all $k \ge k_0$,

(3.41)

$$
\begin{aligned}
&\mathbb{E}\left[\rho^2(e,X_k)\right] \\
&\le \mathbb{E}\left[\rho^2(e,X_{k_0})\right]\exp\left(-C_{p,\mu,K}\sum_{j=k_0+1}^{k} t_j\right) + C(\beta,r,p)\sum_{i=k_0+1}^{k} t_i^2 \exp\left(-C_{p,\mu,K}\sum_{j=i+1}^{k} t_j\right),
\end{aligned}
$$

where the convention $\sum_{j=k+1}^{k} t_j = 0$ is used. With $t_n = \frac{\delta}{n}$, the following inequality holds for all $i \ge k_0$ and $k \ge i$:

$$
\sum_{j=i+1}^{k} t_j = \delta \sum_{j=i+1}^{k}\frac{1}{j} \ge \delta\int_{i+1}^{k+1}\frac{dt}{t} \ge \delta \ln\frac{k+1}{i+1}. \tag{3.42}
$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\rho^2(e,X_k)\right] \\
&\le \mathbb{E}\left[\rho^2(e,X_{k_0})\right]\left(\frac{k_0+1}{k+1}\right)^{\delta C_{p,\mu,K}} + \frac{\delta^2 C(\beta,r,p)}{(k+1)^{\delta C_{p,\mu,K}}}\sum_{i=k_0+1}^{k}\frac{(i+1)^{\delta C_{p,\mu,K}}}{i^2}.
\end{aligned} \tag{3.43}
$$

For $\delta C_{p,\mu,K} > 1$ we have as $k \to \infty$

(3.44)

$$
\frac{\delta^2 C(\beta,r,p)}{(k+1)^{\delta C_{p,\mu,K}}}\sum_{i=k_0+1}^{k}\frac{(i+1)^{\delta C_{p,\mu,K}}}{i^2} \sim \frac{\delta^2 C(\beta,r,p)}{(k+1)^{\delta C_{p,\mu,K}}}\frac{k^{\delta C_{p,\mu,K}-1}}{\delta C_{p,\mu,K}-1} \sim \frac{\delta^2 C(\beta,r,p)}{\delta C_{p,\mu,K}-1}k^{-1}
$$

and

$$
\mathbb{E}\left[\rho^2(e,X_{k_0})\right]\left(\frac{k_0+1}{k+1}\right)^{\delta C_{p,\mu,K}} = o(k^{-1}).
$$

This implies that the sequence $k\mathbb{E}\left[\rho^2(e,X_k)\right]$ is bounded.

Next consider the case $p \ge 2$.

From the proof of Theorem 2.3 we have

$$\mathbb{E}\left[H_p(X_{k+1}) - H_p(e)\right] \le (1 - t_{k+1}C_{p,\mu,K})\mathbb{E}\left[H_p(X_k) - H_p(e)\right] + C(\beta, r, p)t_{k+1}^2 \tag{3.45}$$

which implies

$$\mathbb{E}\left[H_p(X_{k+1}) - H_p(e)\right] \le \mathbb{E}\left[H_p(X_k) - H_p(e)\right]\exp\left(-C_{p,\mu,K}t_{k+1}\right) + C(\beta, r, p)t_{k+1}^2. \tag{3.46}$$

From this, arguing similarly, we obtain that the sequence $k\mathbb{E}\left[H_p(X_k) - H_p(e)\right]$ is bounded. We conclude with (2.1). □

Lemma 3.4. *Assume $\delta > C_{p,\mu,K}^{-1}$ and that H_p is C^2 in a neighbourhood of e. For all $0 < \varepsilon < T$, the sequence of processes $\left(Y_{[nt]}^n\right)_{\varepsilon \le t \le T}$ is tight in $\mathbb{D}([\varepsilon, T], \mathbb{R}^d)$.*

Proof. Denote by $\left(\tilde{Y}_\varepsilon^n = \left(Y_{[nt]}^n\right)_{\varepsilon \le t \le T}\right)_{n\ge1}$, the sequence of processes. We prove that from any subsequence $\left(\tilde{Y}_\varepsilon^{\phi(n)}\right)_{n\ge1}$, we can extract a further subsequence $\left(\tilde{Y}_\varepsilon^{\psi(n)}\right)_{n\ge1}$ that weakly converges in $\mathbb{D}([\varepsilon, 1], \mathbb{R}^d)$.

Let us first prove that $\left(\tilde{Y}_\varepsilon^{\phi(n)}(\varepsilon)\right)_{n\ge1}$ is bounded in L^2.

$$\left\|\tilde{Y}_\varepsilon^{\phi(n)}(\varepsilon)\right\|_2^2 = \frac{[\phi(n)\varepsilon]^2}{\phi(n)}\mathbb{E}\left[\rho^2(e, X_{[\phi(n)\varepsilon]})\right] \le \varepsilon \sup_{n\ge1}\left(n\mathbb{E}\left[\rho^2(e, X_n)\right]\right)$$

and the last term is bounded by lemma 3.3.

Consequently $\left(\tilde{Y}_\varepsilon^{\phi(n)}(\varepsilon)\right)_{n\ge1}$ is tight. So there is a subsequence $\left(\tilde{Y}_\varepsilon^{\psi(n)}(\varepsilon)\right)_{n\ge1}$ that weakly converges in T_eM to the distribution ν_ε. Thanks to Skorohod theorem which allows to realize it as an a.s. convergence and to lemma 3.2 we can apply Theorem 11.2.3 of [20], and we obtain that the sequence of processes $\left(\tilde{Y}_\varepsilon^{\psi(n)}\right)_{n\ge1}$ weakly converges to a diffusion $(y_t)_{\varepsilon\le t\le T}$ with generator $G_\delta(t)$ given by (2.11) and such that y_ε has law ν_ε. This achieves the proof of lemma 3.4. □

Proof of Theorem 2.6. Let $\tilde{Y}^n = \left(Y_{[nt]}^n\right)_{0\le t\le T}$. It is sufficient to prove that any subsequence of $\left(\tilde{Y}^n\right)_{n\ge1}$ has a further subsequence which converges in law to $(y_\delta(t))_{0\le t\le T}$. So let $\left(\tilde{Y}^{\phi(n)}\right)_{n\ge1}$ a subsequence. By lemma 3.4 with $\varepsilon = 1/m$ there exists a subsequence which converges in law on $[1/m, T]$. Then we extract a sequence indexed by m of subsequence and take the diagonal subsequence $\tilde{Y}^{\eta(n)}$. This subsequence converges in $\mathbb{D}((0, T], \mathbb{R}^d)$ to $(y'(t))_{t\in(0,T]}$. On the other hand, as in the proof of lemma 3.4, we have

$$\|\tilde{Y}^{\eta(n)}(t)\|_2^2 \le Ct$$

for some $C > 0$. So $\|\tilde{Y}^{\eta(n)}(t)\|_2^2 \to 0$ as $t \to 0$, which in turn implies $\|y'(t)\|_2^2 \to 0$ as $t \to 0$. The unicity statement in Proposition 2.5 implies that $(y'(t))_{t\in(0,T]}$ and $(y_\delta(t))_{t\in(0,T]}$ are equal in law. This achieves the proof. □

References

[1] B. Afsari, *Riemannian L^p center of mass : existence, uniqueness, and convexity*, Proceedings of the American Mathematical Society, S 0002-9939(2010)10541-5, Article electronically published on August 27, 2010.

[2] M. Arnaudon and X.M. Li, *Barycenters of measures transported by stochastic flows*, The Annals of Probability, 33 (2005), no. 4, 1509–1543

[3] A. Benveniste, M. Goursat and G. Ruget, *Analysis of stochastic approximation schemes with discontinuous and dependent forcing terms with applications to data communication algorithm*, IEEE Transactions on Automatic Control, Vol. AC-25, no. 6, December 1980.

[4] E. Berger, *An almost sure invariance principle for stochastic approximation procedures in linear filtering theory*, the Annals of applied probability, vol. 7, no. 2 (May 1997), pp. 444–459.

[5] M. Emery and G. Mokobodzki, *Sur le barycentre d'une probabilité dans une variété*, Séminaire de Probabilités XXV, Lecture Notes in Mathematics 1485 (Springer, Berlin, 1991), pp. 220–233

[6] S. Gouëzel, *Almost sure invariance principle for dynamical systems by spectral methods* The annals of Probability, Vol 38, no. 4 (2010), pp. 1639–1671

[7] P.T. Fletcher, S. Venkatasubramanian, S. Joshi, *The geometric median on Riemannian manifolds with application to robust atlas estimation*, NeuroImage, 45 (2009), pp. S143–S152

[8] H. Karcher, *Riemannian center of mass and mollifier smoothing*, Communications on Pure and Applied Mathematics, vol XXX (1977), 509–541

[9] R.Z. Khas'Minskii, *On stochastic processes defined by differential equations with a small parameter*, Theory Prob. Appl., vol. XI-2 (1968), pp. 211–228

[10] W.S. Kendall, *Probability, convexity and harmonic maps with small image I: uniqueness and fine existence*, Proc. London Math. Soc. (3) 61 no. 2 (1990) pp. 371–406

[11] W.S. Kendall, *Convexity and the hemisphere*, J. London Math. Soc., (2) 43 no.3 (1991), pp. 567–576

[12] H.W. Kuhn, *A note on Fermat's problem*, Mathematical Programmming, 4 (1973), pp. 98–107

[13] H. Le, *Estimation of Riemannian barycentres*, LMS J. Comput. Math. 7 (2004), pp. 193–200

[14] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Transactions on Automatic Control, Vol. AC-22, no. 4, August 1977.

[15] D. L. MCLeish, *Dependent central limit theorems and invariance principles*, the Annals of Probability, vol. 2, no. 4 (1974), pp. 620–628

[16] A. Nedic and D.P. Bertsekas, *Convergence rate of incremental subgradient algorithms*, Stochastic optimization : algorithms and applications (S. Uryasev and P.M. Pardalos, Editors), Kluwer Academic Publishers (2000), pp. 263–304

[17] L.M.JR. Ostresh, *On the convergence of a class of iterative methods for solving Weber location problem*, Operation. Research, 26 (1978), no. 4

[18] J. Picard, *Barycentres et martingales dans les variétés*, Ann. Inst. H. Poincaré Probab. Statist. 30 (1994), pp. 647–702

[19] A. Sahib, *Espérance d'une variable aléatoire á valeurs dans un espace métrique*, Thése de l'université de Rouen (1998)

[20] D.W. Stroock and S.R.S. Varadhan, *Multidimensional diffusion processes*, Grundlehren der mathematischen Wissenschaften 233, Springer, 1979.

[21] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tohoku Math. J. 43 (1937), pp. 355–386

[22] L.Yang, *Riemannian median and its estimation*, LMS Journal of Computation and Mathematics 13 (2010) pp. 461–479

Laboratoire de Mathématiques et Applications
CNRS: UMR 6086
Université de Poitiers, Téléport 2 - BP 30179
F–86962 Futuroscope Chasseneuil Cedex, France
E-mail address: marc.arnaudon@math.univ-poitiers.fr

Laboratoire de Mathématiques et Applications
CNRS: UMR 6086
Université de Poitiers, Téléport 2 - BP 30179
F–86962 Futuroscope Chasseneuil Cedex, France
E-mail address: clement.dombry@math.univ-poitiers.fr

Laboratoire de Mathématiques et Applications
CNRS: UMR 6086
Université de Poitiers, Téléport 2 - BP 30179
F–86962 Futuroscope Chasseneuil Cedex, France
E-mail address: anthony.phan@math.univ-poitiers.fr

Laboratoire de Mathématiques et Applications
CNRS: UMR 6086
Université de Poitiers, Téléport 2 - BP 30179
F–86962 Futuroscope Chasseneuil Cedex, France
E-mail address: le.yang@math.univ-poitiers.fr