Gaussian distributions on Riemannian symmetric spaces of non-positive curvature

Salem Said, Cyrus Mostajeran, Simon Heuveline

Gaussian distributions on Riemannian symmetric spaces of non-positive curvature

Salem Said (CNRS, Laboratoire LJK – Université Grenoble-Alpes)
Cyrus Mostajeran (Department of Engineering – University of Cambridge)
Simon Heuveline (Centre for Mathematical Sciences – University of Cambridge)

Abstract

This article aims to give a coherent presentation of the theory of Gaussian distributions on Riemannian symmetric spaces and also to report on recent original developments of this theory. The initial goal is to define a family of probability distributions, on any suitable Riemannian manifold, for which maximum-likelihood estimation, based on a finite sequence of observations, is equivalent to computation of the Riemannian barycentre of these observations. As it turns out, this goal is achievable whenever the underlying Riemannian manifold is a Riemannian symmetric space of non-positive curvature. In this case, the required Gaussian distributions are exactly the maximum-entropy distributions, for fixed barycentre and dispersion. The second step is the search for efficient means of computing the normalising factors associated with these distributions. This leads to a fascinating connection with random matrix theory, and even with theoretical physics (Chern-Simons theory), which yields a series of original results that provide exact expressions, as well as high-dimensional asymptotic expansions of the normalising factors. Another outcome of this connection with random matrix theory is the new idea of *duality*, between Gaussian distributions on Riemannian symetric spaces of opposite curvatures. The present article also investigates Bayesian inference for Gaussian distributions on symmetric spaces. This investigation motivates original results regarding Markov-Chain Monte Carlo and convex optimisation on Riemannian manifolds. It also reveals a new open problem (roughly, this concerns the equality of *a posteriori* mode with *a posteriori* barycentre), which should be the focus of future developments.

keywords : Gaussian distribution ; Symmetric space ; Random matrix theory ; Markov-Chain Monte Carlo ; Convex optimisation ; Bayesian inference

1 Introduction

The realisation that an essentially new approach, beyond that of classical statistics, is needed in order to learn from data that live in non-Euclidean spaces, can be credited to Fréchet, who invented what we today call the Fréchet mean, back in 1948 [1].

The Fréchet mean generalises the concept of the mean (average or expectation) of a sequence of observations $(x_1, \ldots, x_N)$, from the classical case where these observations lie in a Euclidean space, to the general setting where they belong to a non-Euclidean space.

Let us call our sample space M (the observations belong to M). If M is a Euclidean space, it has a vector space structure, and the mean of $(x_1, \ldots, x_N)$ is just the arithmetic mean $(x_1 + \ldots + x_N)/N$. If M is a non-Euclidean space, it will have no vector space structure, and this definition will lose all meaning.

To salvage the concept of mean, Fréchet suggested looking at the set of global minima of the sum of squared distances (the factor $1/2$ is included for later convenience)

$$\mathcal{E}(x) = \frac{1}{2}\sum_{n=1}^{N} d^2(x_n, x) \qquad \text{for } x \in M$$

He noted that any global minimum of $\mathcal{E}$ deserves to be called a mean of $(x_1, \ldots, x_N)$. In this way, the mean of a sequence of observations in a non-Euclidean space is well-defined, at the cost of eventually failing to be unique.

Fast-forward to the present, learning from data that live in Riemannian manifolds (a particular class of non-Euclidean spaces) has become central to many applications, ranging from radar signal processing to neuroscience [2, 3], and the Fréchet mean (more descriptively called the Riemannian barycentre) a very popular tool in this respect [4, 5].

Naturally, it also became important to provide a statistical (specifically, inferential) foundation for the use of this tool. One way or another, this lead to the quest for a suitable definition of a Gaussian distribution on a Riemannian manifold. This appeared inevitable, already because of the intimate connection between arithmetic means and Gaussian distributions, in the classical Eucidean case (see Paragraph 2.1, for discussion).

Gaussian distributons, defined as maximum entropy distributions on a Riemannian manifold, for a given Fréchet mean and dispersion, where first introduced by Pennec [6]. For a while, it remained difficult to study these distributions, as there was no practical means of computing the associated normalising factors. However, a first breakthrough came when these factors where expressed as multiple integrals, in the case of Gaussian distributions on the space of real positive-definite matrices [7].

In [8][9], the approach of [7] was generalised to Gaussian distributions on Riemannian symmetric spaces of non-positive curvature, which include hyperbolic spaces, as well as spaces of real, complex and quaternion positive-definite matrices, and spaces of structured (Toeplitz or block-Toeplitz) positive-definite matrices. This opened the way to rigorous learning algorithms for data that live in these spaces (this is partially discussed in [9]).

The introduction of Riemannian symmetric spaces reduced normalising factors of Gaussian distributions to multiple integrals, which could be computed using Monte Carlo techniques [10]. Only very recently, it was realised that the techniques of random matrix theory made it possible to write down both analytic expressions and high-dimensional asymptotic expansions of these multiple integrals. This was studied by the theoretical physics community [11] (see also our paper, currently under review [12]).

The aim of the present article is to give a coherent presentation of the theory of Gaussian distributions on Riemannian symmetric spaces of non-positive curvature, and report on recent original developments of this theory, including (but not limited to) the ones just mentioned. Its main body (Sections 2 and 3) relies on a variety of new results, mostly contained in the habilitation thesis [13] — one advantage of this situation is that the flow of results is not interrupted by their sometimes lengthy proofs, given in [13].

In the following, Section 2 introduced Gaussian distributions and their connection with random matrix theory. Section 3 investigates Bayesian inference of these distributions. Each one of these sections opens with a description of the original results which it contains.

Another, more modest, contribution of this article is Appendix B (not based on [13]). This appendix provides new results on the convergence rates for Riemannian gradient descent, applied to strictly convex and strongly convex functions, defined on a convex subset of a Riemannian manifold. The main results are Propositions B.8 and B.9.

2 Gaussian distributions and RMT

The starting point of this section is a historical discussion of the concept of a Gaussian distribution. This leads up to the definition of Gaussian distributions, adopted in Paragraph 2.2, as a family of distributions $P(\bar{x}, \sigma)$ on a Riemannian manifold M, with parameters $\bar{x} \in M$ and $\sigma > 0$, for which maximum-likelihood estimation of $\bar{x}$ is equivalent to computation of the Riemannian barycentre. It turns out that this definition can be pursued whenever M is a Riemannian symmetric space of non-positive curvature.

Paragraph 2.3 then gives a general expression of the normalising factor $Z(\sigma)$ of the Gaussian distribution $P(\bar{x}, \sigma)$, in the form of a multiple integral (8). When M is a space of positive-definite matrices, or when M is the so-called Siegel domain, (8) is further reduced to a kind of integral familiar in random matrix theory ((10) and (15), respectively).

Paragraph 2.4 states the existence and uniqueness of maximum-likelihood estimates of the parameters $\bar{x}$ and σ. It also states the maximum-entropy property of the Gaussian distribution $P(\bar{x}, \sigma)$, in Proposition 2.5. Paragraph 2.5 provides expressions of the barycentre (shown to be equal to $\bar{x}$) and the covariance tensor of $P(\bar{x}, \sigma)$.

Paragraph 2.6 begins the series of results based on random matrix theory (RMT). These concern Gaussian distributions on the space $\mathrm{H}(N)$ of complex positive-definite matrices. First, the analytic expression of $Z(\sigma)$ is given in Proposition 2.8. Then, an asymptotic expansion of this expression, in the limit where N goes to infinity while $t = N\sigma^2$ remains constant, is given in Proposition 2.9.

Paragraph 2.7 describes the asymptotic distribution of eigenvalues of a random positive-definite matrix in $\mathrm{H}(N)$, drawn from the Gaussian distribution $P(\mathrm{I}_N, \sigma)$ (I_N denotes the $N \times N$ identity matrix). This asymptotic distribution has a probability density function, whose explicit expression is provided in Proposition 2.10.

Paragraph 2.8 introduces Θ distributions. These are classical normal distributions, wrapped around the unitary group $U(N)$, which is the dual symmetric space of $\mathrm{H}(N)$. Proposition 2.11 uncovers an unexpected relationship between Θ distributions on $U(N)$ and Gaussian distributions on $\mathrm{H}(N)$: the normalising factors of these distributions are connected by a simple identity (38).

Proofs of the above-mentioned results may be found in Chapter 3 of [13].

2.1 From Gauss to Shannon

The story of Gaussian distributions is a story of discovery and re-discovery. Different scientists, at different times, were repeatedly lead to these distributions, through different routes. It seems the story began in 1801, on New Year's day, when Giuseppe Piazzi sighted a heavenly body (in fact, the asteroid Ceres), which he thought to be a new planet. Less than six weeks later, this "new planet" disappeared behind the sun. Using a method of least squares, Gauss predicted the area in the sky, where it re-appeared one year later. His justification of this method of least squares (cast in modern language) is that measurement errors follow a family of distributions, which satisfies

Property 1 : maximum-likelihood estimation is equivalent to the least-squares problem.

In his *Theoria motus corporum coelestium (1809)*, he used this property to show that the distribution of measurement errors is (again, in modern language) a Gaussian distribution.

In 1810, Laplace studied the distribution of a quantity, which is the aggregate of a great number of elementary observations. He was lead in this (completely different) way, to the same distribution discovered by Gauss. Laplace was among the first scientists to show

Property 2 : the distribution of the sum of a large number of elementary observations is (asymptotically) a Gaussian distribution.

Around 1860, Maxwell rediscovered Gaussian distributions, through his investigation of the velocity distribution of particles in an ideal gas (which he viewed as freely-colliding perfect elastic spheres). Essentially, he showed that

Property 3 : the distribution of a rotationally-invariant random vector, which has independent components, is a Gaussian distribution.

Kinetic theory lead to another fascinating development, related to Gaussian distributions. Around 1905, Einstein (and, independently, Smoluchowsky) showed that

Property 4 : the distribution of the position of a particle, which is undergoing a Brownian motion, is a Gaussian distribution.

In addition to kinetic theory, alternative routes to Gaussian distributions have been found in quantum mechanics, information theory, and other fields. In quantum mechanics, a Gaussian distribution is a position distribution with minimum uncertainty. That is, it achieves equality in Heisenberg's inequality. In information theory, one may attribute to Shannon the following maximum-entropy characterisation

Property 5 : a probability distribution with maximum entropy, among all distributions with a given mean and variance, is a Gaussian distribution.

The above list of re-discoveries of Gaussian distributions may be extended much longer. However, the main point is the following. In a Euclidean space, identified with $\mathbb{R}^d$, any one of the above five properties leads to the same famous expression of a Gaussian distribution,

$$P(dx|\bar{x},\sigma) = \left(2\pi\sigma^2\right)^{-\frac{d}{2}} \exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right] dx$$

as a probability distribution on $\mathbb{R}^d$, with mean vector $\bar{x} \in \mathbb{R}^d$ and variance parameter $\sigma > 0$ (here dx denotes the Lebesgue measure on $\mathbb{R}^d$).

In non-Euclidean space, each one of these properties may lead to a different distribution, which may then be called a Gaussian distribution, but only from a restricted point

of view. People interested in Brownian motion may call the heat kernel of a Riemannian manifold a Gaussian distribution on that manifold. However, statisticians will not like this definition, since it will (in general) fail to have a straightforward connection to maximum-likelihood estimation.

2.2 The "right" Gaussian

As of now, the following definition of Gaussian distributions is chosen. Gaussian distributions, on a Riemannian manifold M, are a family of distributions $P(\bar{x}, \sigma)$, parameterised by $\bar{x} \in M$ and $\sigma > 0$, such that: a maximum-likelihood estimate $\hat{x}_N$ of $\bar{x}$, based on samples $(x_n\,; n = 1, \ldots, N)$ from $P(\bar{x}, \sigma)$, is a solution of the least-squares problem

$$\text{minimise over } x \in M \quad \mathcal{E}_N(x) = \frac{1}{2}\sum_{n=1}^{N} d^2(x_n, x) \tag{1}$$

This means that $\hat{x}_N$ is an empirical barycentre of the samples (x_n). In order to construct probability distributions $P(\bar{x}, \sigma)$, which satisfy this definition, consider the density profile

$$f(x|\bar{x}, \sigma) \;=\; \exp\left[-\frac{d^2(x, \bar{x})}{2\sigma^2}\right] \tag{2}$$

and the normalising factor,

$$Z(\bar{x}, \sigma) = \int_M f(x|\bar{x}, \sigma)\,\mathrm{vol}(dx) \tag{3}$$

where vol denotes Riemannian volume. If this is finite, then

$$P(dx|\bar{x}, \sigma) \;=\; (Z(\bar{x}, \sigma))^{-1}\, f(x|\bar{x}, \sigma)\,\mathrm{vol}(dx) \tag{4}$$

is a well-defined probability distribution on M. In 2.4, below, it will be seen that $P(\bar{x}, \sigma)$, as defined by (4), is indeed a Gaussian distribution, if M is a Hadamard manifold and also a homogeneous space. The following propositions will then be helpful.

Proposition 2.1. *Let M be a Hadamard manifold, whose sectional curvatures lie in $[\kappa, 0]$, where $\kappa = -c^2$. Then, for any $\bar{x} \in M$ and $\sigma > 0$, if $Z(\bar{x}, \sigma)$ is given by (3),*

$$Z_0(\sigma) \leq\; Z(\bar{x}, \sigma) \;\leq Z_c(\sigma) \tag{5}$$

where $Z_0(\sigma) = (2\pi\sigma^2)^{\frac{d}{2}}$ and $Z_c(\sigma)$ is positive, given by (d denotes the dimension of M)

$$Z_c(\sigma) = \omega_{d-1}\,\frac{\sigma}{(2c)^{d-1}}\sum_{k=0}^{d-1}(-1)^k \binom{d-1}{k} \frac{\Phi\left((d-1-2k)\sigma c\right)}{\Phi'\left((d-1-2k)\sigma c\right)} \tag{6}$$

where ω_{d-1} denotes the area of the unit sphere in $\mathbb{R}^d$, and Φ denotes the standard normal distribution function.

Proposition 2.2. *If M is a Riemannian homogeneous space, and $Z(\bar{x}, \sigma)$ is given by (3), then $Z(\bar{x}, \sigma)$ does not depend on $\bar{x}$. In other words, $Z(\bar{x}, \sigma) = Z(\sigma)$.*

If M is a Hadamard manifold and also a homogeneous space, then both Propositions 2.1 and 2.2 apply to M. Indeed, if M is a Riemannian homogeneous space, then its sectional curvatures lie within a bounded subset of the real line. Therefore, Proposition 2.1 implies $Z(\bar{x}, \sigma)$ is finite for all $\bar{x} \in M$ and $\sigma > 0$. On the other hand, Proposition 2.2 implies that $Z(\bar{x}, \sigma) = Z(\sigma)$. Then, (4) reduces to

$$P(dx|\bar{x}, \sigma) \; = \; (Z(\sigma))^{-1} \exp\left[-\frac{d^2(x, \bar{x})}{2\sigma^2}\right] \mathrm{vol}(dx) \tag{7}$$

and yields a well-defined probability distribution $P(\bar{x}, \sigma)$ on M. This will be the main focus, throughout the following.

Remark : here, readers may wish to recall the concept of a Hadamard manifold, or of a homogeneous space, from [14] (or any other good Riemannian geometry textbook). The point of appealing to these concepts is the following. The assumption that M is a Hadamard manifold implies that geodesic spherical coordinates, which cover all of M, can be introduced at any point $\bar{x} \in M$. Proposition 2.1 is obtained by writing the integral (3) in terms of these spherical coordinates, and then applying Riemannian volume comparison theorems, that state, very roughly speaking, that manifolds with more positive curvature have less volume. On the other hand, to say that M is a homogeneous space means that all points $\bar{x} \in M$ are equivalent, so changing the point of origin $\bar{x}$ does not change the integral (3). This is the key to Proposition 2.2.

2.3 The normalising factor $Z(\sigma)$

Assume now $M = G/K$ is a Riemannian symmetric space which belongs to the non-compact case, described in Appendix A.1. In particular, M is a Hadamard manifold and also a homogeneous space. Thus, for each $\bar{x} \in M$ and $\sigma > 0$, there is a well-defined probability distribution $P(\bar{x}, \sigma)$ on M, given by (7). Here, the normalising factor $Z(\sigma)$ can be expressed as a multiple integral, using the integral formula (74) of Proposition A.1, from Appendix A.1. Applying this proposition (with $o = \bar{x}$), it is enough to note

$$f(\varphi(s, a)|\bar{x}, \sigma) = \exp\left[-\frac{\|a\|_B^2}{2\sigma^2}\right]$$

where $\|a\|_B^2 = B(a, a)$, in terms of the $\mathrm{Ad}(G)$-invariant symmetric bilinear form B (see Appendix A.1). Since this expression only depends on a, (74) yields the following formula

$$Z(\sigma) \; = \; \frac{\omega(S)}{|W|} \int_{\mathfrak{a}} \exp\left[-\frac{\|a\|_B^2}{2\sigma^2}\right] \prod_{\lambda \in \Delta_+} |\sinh \lambda(a)|^{m_\lambda} \, da \tag{8}$$

This formula expresses $Z(\sigma)$ as a multiple integral on the vector space $\mathfrak{a}$. Recall that the dimension of $\mathfrak{a}$ is known as the rank of M [15].

Example 1 : the easiest instance of (8) arises when M is a hyperbolic space of dimension d, and constant sectional curvature equal to -1. Then, M has rank equal to 1, so that $\mathfrak{a} = \mathbb{R}\hat{a}$ for some unit vector $\hat{a} \in \mathfrak{a}$. Since the sectional curvature is equal to -1, there is only one positive root λ, say $\lambda(\hat{a}) = 1$, with multiplicity $m_\lambda = d - 1$. In addition,

$|W| = 2$ because there are two Weyl chambers, $C_+ = \{r\hat{a}\,; r > 0\}$ and $C_- = \{r\hat{a}\,; r < 0\}$. Accordingly, (8) reads

$$Z(\sigma) \;=\; \frac{\omega_{d-1}}{2} \int_{-\infty}^{+\infty} \exp\left[-\frac{r^2}{2\sigma^2}\right] |\sinh(r)|^{d-1}\, dr = \omega_{d-1} \int_0^{+\infty} \exp\left[-\frac{r^2}{2\sigma^2}\right] \sinh^{d-1}(r)\, dr$$

In general, if all distances are divided by $c > 0$, the sectional curvature -1 is replaced by $-c^2$. Thus, when M is a hyperbolic space of dimension d, and sectional curvature $-c^2$,

$$Z(\sigma) = \omega_{d-1} \int_0^{+\infty} \exp\left[-\frac{r^2}{2\sigma^2}\right] (c^{-1}\sinh(cr))^{d-1}\, dr$$

This is exactly $Z_c(\sigma)$, expressed analytically in (6).

Example 2 : another example, also susceptible of analytic expression, is when M is a space of positive-definite matrices with real, complex, or quaternion coefficients. Then, $M = G/K$ with $G = \mathrm{GL}(N, \mathbb{K})$, where $\mathbb{K} = \mathbb{R}, \mathbb{C}$ or $\mathbb{H}$ (real numbers, complex numbers, or quaternions), and $K \subset G$ a maximal compact subgroup, $K = O(N), U(N)$ or $Sp(N)$. In each of these three cases, $\mathfrak{a}$ is the space of $N \times N$ real diagonal matrices, and the positive roots are the linear maps $\lambda(a) = a_{ii} - a_{jj}$ where $i < j$, each one having its multiplicity $m_\lambda = \beta$, ($\beta = 1, 2$ or 4, for $\mathbb{K} = \mathbb{R}, \mathbb{C}$ or $\mathbb{H}$). In addition, $\|a\|_B^2 = 4\mathrm{tr}(a^2)$. The Weyl group W is the groupe of permutation matrices in K, so $|W| = N!$, while $S = K/T_N$ where T_N is the subgroup of all diagonal matrices in K. Replacing all of this into (8), it follows that

$$Z(\sigma) = \frac{\omega_\beta(N)}{N!} \int_{\mathfrak{a}} \prod_{i=1}^{N} \exp\left[-\frac{2a_{ii}^2}{\sigma^2}\right] \prod_{i<j} |\sinh(a_{ii} - a_{jj})|^\beta\, da \tag{9}$$

where $\omega_\beta(N)$ stands for $\omega(S)$, and $da = da_{11} \ldots da_{NN}$. Introducing $x_i = \exp(2a_{ii})$,

$$Z(\sigma) = \frac{\omega_\beta(N)}{2^{NN_\beta} N!} \int_{\mathbb{R}_+^N} |V(x)|^\beta \prod_{i=1}^{N} \rho(x_i, 2\sigma^2) x_i^{-N_\beta}\, dx_i$$

where $N_\beta = (\beta/2)(N-1) + 1$, $\rho(x,k) = \exp(-\log^2(x)/k)$ and $V(x) = \prod_{i<j}(x_j - x_i)$ is the Vandermonde determinant. Finally, using the elementary identity

$$\rho(x,k)x^\alpha = \exp\left[\frac{k}{4}\alpha^2\right] \rho\left(e^{-\frac{k}{2}\alpha}x, k\right)$$

it is immediately found that

$$Z(\sigma) = \frac{\omega_\beta(N)}{2^{NN_\beta} N!} \times \exp\left[-NN_\beta^2(\sigma^2/2)\right] \times \int_{\mathbb{R}_+^N} |V(u)|^\beta \prod_{i=1}^{N} \rho(u_i, 2\sigma^2) du_i \tag{10}$$

For the case $\beta = 2$, the integral in (10) will be expressed analytically in 2.6, below.

Remark : curious readers will want to compute $\omega_\beta(N)$. For example, $\omega_2(N)$ can be found using the Weyl integral formula on $U(N)$ [16]. This yields $\omega_2(N) = \mathrm{vol}(U(N))/(2\pi)^N$. The volume of the unitary group can be found by looking at the normalising factor of a

Gaussian unitary ensemble [17]. Specifically, $\mathrm{vol}(U(N)) = (2\pi)^{(N^2+N)/2}/G(N)$, in terms of $G(N) = \Gamma(1) \times \Gamma(2) \times \ldots \times \Gamma(N)$ (Γ denotes the Euler Gamma function). ■

Example 3 : for this last example, let $M = \mathrm{D}_N$ be the Siegel domain [18]. This is the set of $N \times N$ symmetric complex matrices z, such that $\mathrm{I}_N - z^\dagger z$ is positive-definite. Here, $M = G/K$, where $G \simeq \mathrm{Sp}(N, \mathbb{R})$ (real symplectic group) and $K \simeq U(N)$ (unitary group). Precisely, G is the group of $2N \times 2N$ complex matrices g, with $g^{\mathrm{t}} \Omega g = \Omega$ and $g^\dagger \Gamma g = \Gamma$, where $^{\mathrm{t}}$ denotes the transpose, and where Ω and Γ are the matrices

$$\Omega = \begin{pmatrix} & \mathrm{I}_N \\ -\mathrm{I}_N & \end{pmatrix} \;\; ; \;\; \Gamma = \begin{pmatrix} \mathrm{I}_N & \\ & -\mathrm{I}_N \end{pmatrix}$$

In addition, K is the group of block-diagonal matrices $k = \mathrm{diag}(U, U^*)$ where $U \in U(N)$, and * denotes the conjugate. The action of G on M is given by Möbius transformations,

$$g \cdot z = (Az + B)(Cz + D)^{-1} \quad g = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{11}$$

This action preserves the Siegel metric, which is defined by

$$\langle v,v \rangle_z = \| (\mathrm{I}_N - zz^\dagger)^{-1} v \|_B^2 \qquad \|v\|_B^2 = \frac{1}{2} \mathrm{tr}(vv^\dagger) \tag{12}$$

where each tangent vector v is identified with a symmetric complex matrix. Now [19],

$$\mathfrak{a} = \left\{ \begin{pmatrix} & a \\ a & \end{pmatrix} \; ; \, a = \mathrm{diag}(a_{11}, \ldots, a_{NN}) \right\} \tag{13}$$

The positive roots are $\lambda(a) = a_{ii} - a_{jj}$ for $i < j$, and $\lambda(a) = a_{ii} + a_{jj}$ for $i \leq j$, all with $m_\lambda = 1$. The order of the Weyl group is $|W| = 2^N N!$, and $\omega(S) = \mathrm{vol}(U(N))/2^N$. Replacing into (8), it follows that

$$Z(\sigma) \;=\; \frac{\mathrm{vol}(U(N))}{2^{2N} N!} \int_{\mathfrak{a}} \prod_{i=1}^{N} \exp\left[-\frac{a_{ii}^2}{2\sigma^2} \right] \prod_{i<j} \sinh |a_{ii} - a_{jj}| \prod_{i \leq j} \sinh |a_{ii} + a_{jj}| \, da \tag{14}$$

or, after introducing $u_i = \cosh(2a_{ii})$,

$$Z(\sigma) \;=\; 2^{-2N} \mathrm{vol}(U(N)) \times \int_{\mathrm{C}^N} V(u) \prod_{n=1}^{N} w(u_i, 8\sigma^2) du_i \tag{15}$$

where $\mathrm{C}^N = \{ u \in \mathbb{R}_+^N : u_1 \leq u_2 \leq u_1 \leq \ldots \leq u_N \}$ and $w(u,k) = \exp(-\mathrm{acosh}^2(u)/k)$, while $V(u)$ is the Vandermonde determinant, as in (10).

2.4 MLE and maximum entropy

Let M be a Hadamard manifold, which is also a homogeneous space. Consider the family of distributions $P(\bar{x}, \sigma)$ on M, given by (7) for $\bar{x} \in M$ and $\sigma > 0$. This family of distributions fits the definition of Gaussian distributions, stated at the beginning of 2.2.

Proposition 2.3. *Let $P(\bar{x}, \sigma)$ be given by (7), for $\bar{x} \in M$ and $\sigma > 0$. The maximum-likelihood estimate of the parameter $\bar{x}$, based on samples $(x_n\,; n = 1, \ldots, N)$ from $P(\bar{x}, \sigma)$, is unique and equal to the empirical barycentre $\hat{x}_N$ of the samples (x_n).*

This proposition is almost immediate. From (7), one has the log-likelihood function

$$\ell(\bar{x}, \sigma) \;=\; -N \log Z(\sigma) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} d^2(x_n, \bar{x}) \tag{16}$$

Since the first term does not depend on $\bar{x}$, one may maximise $\ell(\bar{x}, \sigma)$, first over $\bar{x}$ and then over σ. Clearly, maximising over $\bar{x}$ is equivalent to minimising the sum of squared distances $d^2(x_n, \bar{x})$. This is just the least-squares problem (1), whose solution is the empirical barycentre $\hat{x}_N$. Moreover, $\hat{x}_N$ is unique, since M is a Hadamard manifold [20][4].

Consider now maximum-likelihood estimation of σ. This is better carried out in terms of the natural parameter $\eta = (-2\sigma^2)^{-1}$, or in terms of the moment parameter $\delta = \psi'(\eta)$, where $\psi(\eta) = \log Z(\sigma)$ and the prime denotes the derivative.

Proposition 2.4. *The function $\psi(\eta)$, just defined, is a strictly convex function, which maps the half-line $(-\infty, 0)$ onto $\mathbb{R}$. The maximum-likelihood estimates of the parameters η and δ are*

$$\hat{\eta}_N = (\psi')^{-1}(\hat{\delta}_N) \quad \textit{and} \quad \hat{\delta}_N = \frac{1}{N} \sum_{n=1}^{N} d^2(x_n, \hat{x}_N) \tag{17}$$

where $(\psi')^{-1}$ denotes the reciprocal function.

Remark : $\hat{\eta}_N$ in (17) is well-defined, since the range of ψ' is equal to $(0, \infty)$. Indeed, one has the following inequalities, analogous to (5),

$$\psi_0'(\eta) \;\leq \psi'(\eta) \leq\; \psi_c'(\eta) \tag{18}$$

where $\psi_0(\eta) = \log Z_0(\sigma)$, and $\psi_c(\eta) = \log Z_c(\sigma)$. Now, $\psi_0'(\eta) = n\sigma^2$, which increases to $+\infty$ when σ increases to $+\infty$. On the other hand, since $\eta = (-2\sigma^2)^{-1}$,

$$\psi_c'(\eta) \;=\; \sigma^3 \frac{d}{d\sigma}(\log Z_c(\sigma)) \tag{19}$$

which, from (6), is $= 0$ when $\sigma = 0$. Thus, it follows from (18) that ψ' maps the half-line $(-\infty, 0)$ onto the half-line $(0, +\infty)$. ■

An alternative definition of Gaussian distributions is provided by their maximum-entropy property, stated in the following proposition. Here, entropy specifically means Shannon's differential entropy. If P is a probability distribution on M, with probability density function p, this entropy is equal to

$$S(P) = \int_M (\log p(x))\, p(x) \mathrm{vol}(dx)$$

Proposition 2.5. *The Gaussian distribution $P(\bar{x}, \sigma)$ is the unique distribution on M, having maximum Shannon entropy, among all distributions P with given barycentre $\bar{x}$ and dispersion $\delta = \mathbb{E}_{x \sim P}[d^2(x, \bar{x})]$. Its entropy is equal to $\psi^*(\delta)$ where ψ^* is the Legendre transform of ψ.*

2.5 Barycentre and covariance

Let M be a Hadamard manifold, which is also a homogeneous space. Consider the barycentre and covariance of the Gaussian distribution $P(\bar{x}, \sigma)$ on M, given by (7).

First, it should be noted $P(\bar{x}, \sigma)$ has a well-defined Riemannian barycentre, since it has finite second-order moments. To see that this is true, it is enough to note that

$$\int_M d^2(\bar{x}, x)\, P(dx|\bar{x}, \sigma) \,<\, \infty$$

Ineded, this integral is just $\psi'(\eta)$ in (18).

Proposition 2.6. *Let $P(\bar{x}, \sigma)$ be given by (7), for $\bar{x} \in M$ and $\sigma > 0$. The Riemannian barycentre of $P(\bar{x}, \sigma)$ is equal to $\bar{x}$.*

The proof of this proposition relies on the fact that the so-called variance function

$$\mathcal{E}(x) \,=\, \frac{1}{2} \int_M d^2(x, y)\, P(dy|\bar{x}, \sigma) \tag{20}$$

is strongly convex (see [13], Paragraph 2.2.3). Thus, if $\mathrm{grad}\mathcal{E}(\bar{x}) = 0$, then $\bar{x}$ is the global minimiser of $\mathcal{E}$, and therefore the barycentre of $P(\bar{x}, \sigma)$. However, $\mathrm{grad}\mathcal{E}(\bar{x}) = 0$ follows by a direct application of the following "Fisher's identity",

$$\int_M \left(\mathrm{grad}_{\bar{x}} \log p(x|\bar{x}, \sigma)\right) P(dx|\bar{x}, \sigma) = 0$$

where $\mathrm{grad}_{\bar{x}}$ denotes the gradient with respect to $\bar{x}$, defined according to the Riemannian metric of M, and $p(x|\bar{x}, \sigma)$ is the probability density function, appearing in (7).

The covariance form of $P(\bar{x}, \sigma)$ is the symmetric bilinear form $C_{\bar{x}}$ on $T_{\bar{x}}M$,

$$C_{\bar{x}}(u, v) \,=\, \int_M \langle u, \mathrm{Exp}_{\bar{x}}^{-1}(x)\rangle \,\langle \mathrm{Exp}_{\bar{x}}^{-1}(x), v\rangle \, p(x|\bar{x}, \sigma)\mathrm{vol}(dx) \qquad u\,, v \in T_{\bar{x}}M \tag{21}$$

where Exp denotes the Riemannian exponential map (Exp^{-1} is well-defined, since M is a Hadamard manifold).

With $\sigma > 0$ fixed, the map which assigns to $\bar{x} \in M$ the covariance form $C_{\bar{x}}$ is a (0,2)-tensor field on M, here called the covariance tensor of $P(\bar{x}, \sigma)$. In order to compute this tensor field, consider the following situation.

Assume $M = G/K$ is a Riemannian symmetric space. Here, $K = K_o$, the stabiliser in G of $o \in M$. For $k \in K$ and $u \in T_oM$, it is clear $k{\cdot}u \in T_oM$. This defines a representation of K in the tangent space T_oM, called the isotropy representation. One says that M is an irreducible symmetric space, if this isotropy representation is irreducible.

If M is not irreducible, then it is a product of irreducible Riemannian symmetric spaces $M = M_1 \times \ldots \times M_s$ [15] (Proposition 5.5, Chapter VIII. This is the de Rham decomposition of M). Accordingly, for $x \in M$ and $u \in T_xM$, one may write $x = (x_1, \ldots, x_s)$ and $u = (u_1, \ldots, u_s)$, where $x_r \in M_r$ and $u_r \in T_{x_r}M_r$. Now, looking back at (7), it may be seen that

$$p(x|\bar{x}, \sigma) \,=\, \prod_{r=1}^{s} p(x_r|\bar{x}_r, \sigma) \qquad p(x_r|\bar{x}_r, \sigma) \,=\, (Z_r(\sigma))^{-1} \exp\left[-\frac{d^2(x_r, \bar{x}_r)}{2\sigma^2}\right] \tag{22}$$

For the following proposition, let $\eta = (-2\sigma^2)^{-1}$ and $\psi_r(\eta) = \log Z_r(\sigma)$.

Proposition 2.7. *Assume that M is a product of irreducible Riemannian symmetric spaces, $M = M_1 \times \ldots \times M_s$. The covariance tensor C in (21) is given by*

$$C_{\bar{x}}(u,u) \;=\; \sum_{r=1}^{s} \frac{\psi_r'(\eta)}{\dim M_r} \|u_r\|^2_{\bar{x}_r} \tag{23}$$

for $u \in T_{\bar{x}}M$ where $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_s)$ and $u = (u_1, \ldots, u_s)$, with $\bar{x}_r \in M_r$ and $u_r \in T_{\bar{x}_r}M_r$.

Example: let $M = \mathrm{H}(N)$, the space of $N \times N$ Hermitian positive-definite matrices, so $M = \mathrm{GL}(N,\mathbb{C})/U(N)$, with $U(N)$ the stabiliser of $o = \mathrm{I}_N$ ($N \times N$ identity matrix). The de Rham decomposition of M is $M = M_1 \times M_2$, where $M_1 = \mathbb{R}$ and M_2 is the submanifold whose elements are those $x \in M$ such that $\det(x) = 1$. Accordingly, each $\bar{x} \in M$ is identified with the couple $(\bar{x}_1, \bar{x}_2)$,

$$\bar{x}_1 \;=\; \frac{1}{N}\log\det(\bar{x}) \quad \bar{x}_2 \;=\; (\det(\bar{x}))^{-1/N}\,\bar{x}$$

and each $u \in T_{\bar{x}}M$ is written $u = u_1\bar{x} + u_2$

$$u_1 \;=\; \frac{1}{N}\mathrm{tr}(\bar{x}^{-1}u) \quad u_2 \;=\; u - \frac{1}{N}\mathrm{tr}(\bar{x}^{-1}u)\,\bar{x}$$

These may be replaced into expression (23),

$$C_{\bar{x}}(u,u) \;=\; \psi_1'(\eta)\,u_1^2 \;+\; \frac{\psi_2'(\eta)}{N^2-1}\|u_2\|^2_{\bar{x}_2} \tag{24}$$

where $\psi_1(\eta) = \log\,(2\pi\sigma^2)^{\frac{1}{2}}$, and $\psi_2(\eta) = \log Z(\sigma) - \psi_1(\eta)$ ($Z(\sigma)$ is given by (26) in 2.6, below). After a direct calculation, this can be brought under the form

$$C_{\bar{x}}(u,u) \;=\; g_1(\sigma)\mathrm{tr}^2(\bar{x}^{-1}u) + g_2(\sigma)\mathrm{tr}(\bar{x}^{-1}u)^2 \tag{25}$$

where $g_1(\sigma)$ and $g_2(\sigma)$ are certain functions of σ.

Remark: as a corollary of Proposition 2.7, the covariance tensor C is a G-invariant Riemannian metric on M. This is clear, for example, in the special case of (25), which coincides with the general expression of a $\mathrm{GL}(N,\mathbb{C})$-invariant metric. ■

2.6 $Z(\sigma)$ from RMT

Random matrix theory is very helpful in the calculation of integrals such as (10) and (15), leading both to exact expressions and to asymptotic expansions of these integrals. Here, this is illustrated for the integral (10), with $\beta = 2$. This corresponds to $M = \mathrm{H}(N)$, the space of $N \times N$ Hermitian positive-definite matrices. In this case, it is possible to provide an analytic formula for the normalising factor $Z(\sigma)$.

Proposition 2.8. *When $M = \mathrm{H}(N)$, the normalising factor $Z(\sigma)$, given by (10) with $\beta = 2$, admits of the following analytic expression*

$$Z(\sigma) = \frac{\omega_2(N)}{2^{N^2}}\left(2\pi\sigma^2\right)^{\frac{N}{2}} \exp\left[\left(\frac{N^3-N}{6}\right)\sigma^2\right] \prod_{n=1}^{N-1}\left(1-e^{-n\sigma^2}\right)^{N-n} \tag{26}$$

The proof of this proposition is a direct application of a well-known formula from random matrix theory [17]. The integral in (10) reads

$$I_N(\sigma) = \int_{\mathbb{R}_+^N} |V(u)|^\beta \prod_{i=1}^N \rho(u_i, 2\sigma^2) du_i$$

According to [17] (Chapter 5, Page 79), if $(p_n\,; n = 0, 1, \ldots)$ are orthonormal polynomials with respect to the weight function $\rho(u, 2\sigma^2)$ on $\mathbb{R}_+$, then $I_N(\sigma)$ is given by

$$I_N(\sigma) \;=\; N! \prod_{n=0}^{N-1} p_{nn}^{-2} \tag{27}$$

where p_{nn} is the leading coefficient in p_n. The required orthonormal polynomials p_n are given by $p_n = (2\pi\sigma^2)^{-\frac{1}{4}} s_n$, where s_n are the Stieltjes-Wigert polynomials [21] (Page 33). Looking up the expression of these polynomials, it is easy to find

$$p_{nn}^{-2} \;=\; \left(2\pi\sigma^2\right)^{\frac{1}{2}} \exp\left[\frac{(2n+1)^2}{2}\sigma^2\right] \prod_{m=1}^{n} \left(1 - e^{-m\sigma^2}\right)$$

Then, working out the product (27) and replacing into (10), one is lead to (26).

Moving on, it is possible to derive an asymptotic expression of $Z(\sigma)$, valid in the limit where N goes to infinity while the product $t = N\sigma^2$ remains constant.

Proposition 2.9. *Let $Z(\sigma)$ be given by (26). If $N \to \infty$, while $t = N\sigma^2$ remains constant, then the following equivalence holds,*

$$\frac{1}{N^2} \log Z(\sigma) \sim -\frac{1}{2} \log\left(\frac{2N}{\pi}\right) + \frac{3}{4} + \frac{t}{6} - \frac{\mathrm{Li}_3(e^{-t}) - \zeta(3)}{t^2} \tag{28}$$

where $\mathrm{Li}_3(x) = \sum_{k=1}^\infty x^k/k^3$ for $|x| < 1$ (the trilogarithm), and ζ is the Riemann Zeta function.

The main idea behind (28) is that, taking the logarithm in (26), the product on the right-hand side turns into a Riemann sum for the improper integral

$$\int_0^1 (1-x) \log\left(1 - e^{-tx}\right) dx = -(\mathrm{Li}_3(e^{-t}) - \zeta(3))/t^2$$

where the equality follows by integrating term-by-term the power series of the logarithm.

2.7 The asymptotic distribution

From the point of view of random matrix theory, a Gaussian distribution $P(\mathrm{I}_N, \sigma)$ on $M = \mathrm{H}(N)$ defines a unitary matrix ensemble. If x is a random matrix, drawn from this ensemble, and $(x_i\,; i = 1, \ldots, N)$ are its eigenvalues, which all belong to $(0, \infty)$, then the empirical distribution ν_N, which is given by (as usual, δ_{x_i} is the Dirac distribution at x_i)

$$\nu_N(B) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \delta_{x_i}(B)\right] \tag{29}$$

for measurable $B \subset (0, \infty)$, converges to an absolutely continuous distribution ν_t, when N goes to infinity, while the product $t = N\sigma^2$ remains constant.

Proposition 2.10. *Let $c = e^{-t}$ and $a(t) = c(1+\sqrt{1-c})^{-2}$ while $b(t) = c(1-\sqrt{1-c})^{-2}$. When N goes to infinity, while the product $t = N\sigma^2$ remains constant, the empirical distribution ν_N converges weakly to the distribution ν_t with probability density function*

$$\frac{d\nu_t}{dx}(x) = \frac{1}{\pi t x} \arctan\left(\frac{4e^t x - (x+1)^2}{x+1}\right) \mathbf{1}_{[a(t),b(t)]}(x) \tag{30}$$

where $\mathbf{1}_{[a(t),b(t)]}$ denotes the indicator function of the interval $[a(t), b(t)]$.

Remark : as one should expect, when $t = 0$ (so $\sigma^2 = 0$), $a(t) = b(t) = 1$. ■

The proof of Proposition 2.10 is a relatively direct application of a result in [22] (Page 191). In fact, the intergration variables in (10) are $u_i = e^t x_i$. Let $\tilde{\nu}_N$ be the empirical distribution of the u_i (this is the same as (29), but with u_i instead of x_i). By applying [17] (Chapter 5, Page 81),

$$\tilde{\nu}_N(B) = \frac{1}{N} \int_B R_N^{(1)}(u)(du) \tag{31}$$

for measurable $B \subset (0, \infty)$, where the one-point correlation function $R_N^{(1)}(u)$ is given by

$$R_N^{(1)}(u) = \rho(u, 2\sigma^2) \sum_{n=0}^{N-1} p_n^2(u) \tag{32}$$

in the notation of 2.6 (p_n are orthonormal polynomials, with respect to the weight $\rho(u, 2\sigma^2)$). According to [23] (Page 133), $\tilde{\nu}_N$ given by (31) converges weakly to the so-called equilibrium distribution $\tilde{\nu}_t$, which minimises the electrostatic energy functional

$$E(\nu) = \frac{1}{t} \int_0^\infty \frac{1}{2} \log^2(u) \nu(du) - \int_0^\infty \int_0^\infty \log|u - v| \nu(du)\nu(dv) \tag{33}$$

over probability distributions ν on $(0, \infty)$. Also according to [23] (Page 133), this equilibrium distribution is the asymptotic distribution of the zeros of the polynomial p_N (in the limit $N \to \infty$ while $N\sigma^2 = t$). Fortunately, p_N is just a constant multiple of the Stieltjes-Wigert polynomial s_N [21] (Page 33). Therefore, the required asymptotic distribution of zeros can be read from [22] (Page 191). Finally, (30) follows by introducing the change of variables $x = e^{-t}u$.

Remark : in [24], the equilibrium distribution $\tilde{\nu}_t$ is derived directly, by searching for stationary distributions of the energy functional (33). This leads to a singular integral equation, whose solution reduces to a Riemann-Hilbert problem. Astoundingly, the Gaussian distributions on H(N), as introduced in the present chapter, provide a matrix model for Chern-Simons quantum field theory (a detailed account is given in [24]). ■

2.8 Duality : the Θ distributions

Recall the Riemannian symmetric space $M = \mathrm{H}(N)$ of 2.6. Its dual space is the unitary group $M^* = U(N)$ (the definition of *duality* may be found in Appendix A.2).

Consider now a family of distributions on M^*, which will be called Θ distributions, and which display an interesting connection with Gaussian distributions on M, studied

in 2.6. Recall Jacobi's ϑ function[1],

$$\vartheta(e^{i\phi}|\sigma^2) \;=\; \sum_{m=-\infty}^{+\infty} \exp(-m^2\sigma^2 + 2mi\phi)$$

As a function of ϕ, up to some minor modifications, this is a wrapped normal distribution (in other words, the heat kernel of the unit circle),

$$\frac{1}{2\pi}\vartheta(e^{i\phi}|\tfrac{\sigma^2}{2}) \;=\; \sum_{m=-\infty}^{\infty} \exp\left[-\frac{(2\phi - 2m\pi)^2}{2\sigma^2}\right]$$

Each $x \in M^*$ can be written $x = k{\cdot}e^{i\theta}$ where $k \in U(N)$ and $e^{i\theta} = \mathrm{diag}(e^{i\theta_j}\,; j = 1, \ldots, N)$. Here, $k \cdot y = kyk^\dagger$ for $y \in M^*$. With this notation, define the following matrix ϑ function,

$$\Theta\left(x\middle|\sigma^2\right) \;=\; k \cdot \vartheta(e^{i\theta}|\tfrac{\sigma^2}{2}) \tag{34}$$

which is obtained from x by applying Jacobi's ϑ function to each eigenvalue of x. Further, consider the positive function,

$$f_*(x|\bar{x},\sigma) \;=\; \det\left[\left(2\pi\sigma^2\right)^{\frac{1}{2}}\,\Theta\big(x\bar{x}^\dagger\big|\sigma^2\big)\right] \tag{35}$$

where $\bar{x} \in M^*$. This is also equal to

$$\det\left[\left(2\pi\sigma^2\right)^{\frac{1}{2}}\,\Theta\big(\bar{x}^\dagger x\big|\sigma^2\big)\right]$$

since the matrices $x\bar{x}^\dagger$ and $\bar{x}^\dagger x$ are similar. Then, let $Z_{M^*}(\sigma)$ denote the normalising constant

$$Z_{M^*}(\sigma) = \int_{M^*} f_*(x|\bar{x},\sigma)\,\mathrm{vol}(dx) \tag{36}$$

which does not depend on $\bar{x}$, as can be seen, by introducing the new variable of integration $z = x\bar{x}^\dagger$, and using the invariance of $\mathrm{vol}(dx)$.

Now, define a Θ distribution $\Theta(\bar{x},\sigma)$ as the probability distribution on M^*, whose probability density function, with respect to $\mathrm{vol}(dx)$, is given by

$$p_*(x|\bar{x},\sigma) \;=\; \left(Z_{M^*}(\sigma)\right)^{-1} f_*(x|\bar{x},\sigma) \tag{37}$$

Proposition 2.11. *Let $Z_M(\sigma) = Z(\sigma)$ be given by (26), and $Z_{M^*}(\sigma)$ be given by (36). Then, the following equality holds*

$$\frac{Z_M(\sigma)}{Z_{M^*}(\sigma)} = \exp\left[\left(\frac{N^3 - N}{6}\right)\sigma^2\right] \tag{38}$$

Remark : the Gaussian density (7) on M, and the Θ distribution density (37) on M^* are apparently unrelated. Therefore, it is interesting to note their normalising constants $Z_M(\sigma)$ and $Z_{M^*}(\sigma)$ scale together according to the simple relation (38). The connection between the two distributions is due to the duality between M and M^*. ■

[1]To follow the original notation of Jacobi [25], this should be written $\vartheta(e^{i\phi}|q)$ where $q = e^{-\sigma^2}$. In other popular notations, this function is called ϑ_{00} or ϑ_3.

3 Gaussian distriutions and Bayesian inference

This section aims to investigate Bayesian inference for Gaussian distributions. Precisely, it aims to study Bayesian estimation of the parameter x of a Gaussian distribution $P(x, \sigma)$, when this parameter is assigned a prior density which is also Gaussian, say $P(z, \tau)$. Both the prior density $P(z, \tau)$ and the likelihood density $P(x, \sigma)$ are defined on a Riemannian symmetric space of non-positive curvature M [2].

Paragraph 3.1 begins by expressing the posterior density $\pi(x)$, based on the general definition

$$\pi(x) \propto \text{ prior density } \times \text{ likelihood density}$$

Here, $\pi(x)$ will remain partially unknown, as the missing normalising factor cannot be determined. Then, two Bayesian estimators are studied. The maximum *a posteriori* $\hat{x}_{\text{MAP}}$ is the mode of $\pi(x)$,

$$\hat{x}_{\text{MAP}} = \text{argmax}_{x \in M} \, \pi(x)$$

while the minimum mean square error estimator $\hat{x}_{\text{MMS}}$, classically understood as the mean (expectation) of the posterior density, is here the Riemanian barycentre of $\pi(x)$.

It is seen that $\hat{x}_{\text{MAP}}$ can be computed directly, being a geodesic convex combination of the prior barycentre z and a new observation y, with respective weights $1 - \rho$ and ρ, where $\rho = \tau^2/(\sigma^2 + \tau^2)$. On the other hand, $\hat{x}_{\text{MMS}}$ seems much harder to compute.

However, Proposition 3.1 states that $\hat{x}_{\text{MMS}} = \hat{x}_{\text{MAP}}$ if $\rho = 1/2$, and Proposition 3.2 states that, in the special case where M is a hyperbolic space, $\hat{x}_{\text{MMS}}$ is a geodesic convex combination of z and y, just like $\hat{x}_{\text{MAP}}$, but with different weights, say $(1 - t^*)$ and t^*.

Paragraph 3.2 reports on numerical experiments which show, again in the special case where M is a hyperbolic space, that $\hat{x}_{\text{MMS}}$ and $\hat{x}_{\text{MAP}}$ lie very close to each other, and that they even appear to be equal (this would mean $t^* = \rho$). At present, I am unaware of any mathematical explanation of this phenomenon.

Paragraph 3.3 describes the computational tools employed in calculating $\hat{x}_{\text{MMS}}$. First, Proposition 3.3 provides easy-to-verify sufficient conditions, for the geometric ergodicity of an isotropic Metropolis-Hastings Markov chain, in a Riemannian symmetric space M. These conditions are shown to apply in the case of the posterior density $\pi(x)$, making it possible to generate geometrically ergodic samples $(x_n \, ; n \geq 1)$ from this density.

The next Proposition 3.4 states that the empirical barycentre $\bar{x}_N$ of the samples $(x_1, \ldots, x_N)$ converges almost-surely to $\hat{x}_{\text{MMS}}$, so $\bar{x}_N$ may be used to approximate $\hat{x}_{\text{MMS}}$ to any required accuracy.

Concretely, computing the empirical barycentre $\bar{x}_N$ requires solving a strongly convex optimisation problem on the Riemannian manifold M (here, convexity is with respect to the Riemannian connection of M [26]. This has come to be called "geodesic convexity"). Appendix B is devoted to a brief but systematic study of convex optimisation on Riemannian manifolds. Specifically, it establishes the rate of convergence of Riemannian gradient descent schemes, applied to strictly convex or strongly convex cost functions.

The gradient descent schemes under consideration are retraction schemes (not limited to the Riemannian exponential) with a constant step-size. The problem is then to find the largest possible step-size which guarantees a certain rate of convergence. Proposition B.8 addresses this problem for strictly convex functions, and Proposition B.9 for strongly

[2] Proofs of the results stated in this section can be found in Chapter 4 of [13].

convex functions. For any strictly convex cost function, and suitable retraction, Proposition B.8 gives the largest possible step-size which guarantees a rate of convergence at least as fast as $O(1/t)$ (t is the number of iterations). Proposition B.9 does the same for strongly convex functions, but with an exponential rate of convergence. In fact, with regard to the original motivation of computing $\bar{x}_N$, this ensures that a gradient descent scheme, using the Riemannian exponential, converges after only a few iterations.

3.1 MAP versus MMS

Assume $M = G/K$ is a Riemannian symmetric space which belongs to the non-compact case, described in Appendix A.1. Recall the Gaussian distribution $P(x, \sigma)$ on M given by its probability density function (7)

$$p(y|x, \sigma) \;=\; (Z(\sigma))^{-1} \exp\left[-\frac{d^2(y, x)}{2\sigma^2}\right] \tag{39}$$

In 2.4, it was seen that maximum-likelihood estimation of the parameter x, based on samples $(y_n\,; n \;=\; 1, \ldots, N)$, amounts to computing the empirical barycentre of these samples. The one-sample maximum-likelihood estimate, given a single observation y, is therefore $\hat{x}_{ML} = y$.

Instead of maximum-likelihood estimation, consider Bayesian estimation of x, based on the observation y. To do so, assign to x a prior density, which is also Gaussian,

$$p(x|z, \tau) \;=\; (Z(\tau))^{-1} \exp\left[-\frac{d^2(x, z)}{2\tau^2}\right] \tag{40}$$

Upon observation of y, Bayesian inference concerning x is carried out using the posterior density

$$\pi(x) \propto \exp\left[-\frac{d^2(y, x)}{2\sigma^2} - \frac{d^2(x, z)}{2\tau^2}\right] \tag{41}$$

where $\propto$ indicates a missing (unknown) normalising factor.

In particular, the maximum *a posteriori* estimator $\hat{x}_{\text{MAP}}$ of x is equal to the mode of the posterior density $\pi(x)$. In other words, $\hat{x}_{\text{MAP}}$ minimises the weighted sum of squared distances $d^2(y, x)/\sigma^2 + d^2(x, z)/\tau^2$. This is expressed in the following notation[3],

$$\hat{x}_{\text{MAP}} \;=\; z \,\#_\rho\, y \quad \text{where } \rho = \frac{\tau^2}{\sigma^2 + \tau^2} \tag{42}$$

Thus, $\hat{x}_{\text{MAP}}$ is a geodesic convex combination of the prior barycentre z and the observation y, with respective weights $\sigma^2/(\sigma^2 + \tau^2)$ and $\tau^2/(\sigma^2 + \tau^2)$.

On the other hand, the minimum mean square error estimator $\hat{x}_{\text{MMS}}$ is the barycentre of the posterior density $\pi(x)$. That is, $\hat{x}_{\text{MMS}}$ is the unique global minimiser of

$$\mathcal{E}_\pi(w) \;=\; \frac{1}{2}\int_M d^2(w, x)\pi(x)\,\text{vol}(dx) \tag{43}$$

[3]If $p, q \in M$ and $c : [0, 1] \to M$ is a geodesic curve with $c(0) = p$ and $c(1) = q$, then $p \,\#_t\, q = c(t)$, for $t \in [0, 1]$. In other words, $p \,\#_t\, q$ is a geodesic convex combination of p and q, with respective weights $(1 - t)$ and t.

While it is easy to compute $\hat{x}_{\text{MAP}}$ from (42), it is much harder to find $\hat{x}_{\text{MMS}}$, as this requires minimising the integral (43), where the density $\pi(x)$ is known only up to normalisation.

Still, there is one special case where these two estimators are equal.

Proposition 3.1. *In the above notation, if $\sigma^2 = \tau^2$ (that is $\rho = 1/2$), then $\hat{x}_{\text{MMS}} = \hat{x}_{\text{MAP}}$.*

When M is a Euclidean space, it is well-known that $\hat{x}_{\text{MMS}} = \hat{x}_{\text{MAP}}$ for any value of ρ. When M is a space of constant negative curvature, the following proposition indicates $\hat{x}_{\text{MMS}}$ and $\hat{x}_{\text{MAP}}$ cannot be too far away from one another.

Proposition 3.2. *In the above notation, if M is a space of constant negative curvature (hyperbolic space), then $\hat{x}_{\text{MMS}} = z \#_{t^*} y$ for some $t^* \in (0, 1)$.*

3.2 Bounding the distance

In general, one expects $\hat{x}_{\text{MMS}}$ and $\hat{x}_{\text{MAP}}$ to be different from one another, when $\rho \neq 1/2$. However, when M is a space of constant negative curvature, Proposition 3.2 shows the distace between these two estimators is always less than the distance between z and y.

Surprisingly (again when M is a space of constant negative curvature), numerical experiments show that $\hat{x}_{\text{MMS}}$ and $\hat{x}_{\text{MAP}}$ lie very close to each other, and that they even appear to be equal. I am unaware of any mathematical explanation of this phenomenon.

It is possible to bound the distance between $\hat{x}_{\text{MMS}}$ and $\hat{x}_{\text{MAP}}$, using the fundamental contraction property [20] (this is an immediate application of Jensen's inequality, as explained in the proof of Theorem 6.3 in [20]).

$$d(\hat{x}_{\text{MMS}}, \hat{x}_{\text{MAP}}) \leq W(\pi, \delta_{\hat{x}_{\text{MAP}}}) \tag{44}$$

where W denotes the Kantorovich (L^1-Wasserstein) distance, and $\delta_{\hat{x}_{\text{MAP}}}$ denotes the Dirac probability distribution concentrated at $\hat{x}_{\text{MAP}}$. Now, the right-hand side of (44) is equal to the first-order moment

$$m_1(\hat{x}_{\text{MAP}}) = \int_M d(\hat{x}_{\text{MAP}}, x)\pi(x)\text{vol}(dx) \tag{45}$$

Of course, the upper bound in (44) is not tight, since it is strictly positive, even when $\rho = 1/2$, as one may see from (45).

It will be shown below that a Metropolis-Hastings algorithm, with Gaussian proposals, can be used to generate geometrically ergodic samples $(x_n\,; n \geq 1)$ from the posterior density π. It is therefore possible to approximate (45) by an empirical average

$$\bar{m}_1(\hat{x}_{\text{MAP}}) = \frac{1}{N}\sum_{n=1}^{N} d(\hat{x}_{\text{MAP}}, x_n) \tag{46}$$

In addition, the samples (x_n) can be used to compute a convergent approximation of $\hat{x}_{\text{MMS}}$. Precisely, the empirical barycentre $\bar{x}_{\text{MMS}}$ of the samples $(x_1, \ldots, x_N)$ converges almost-surely to $\hat{x}_{\text{MMS}}$ (this is a result of Proposition 3.4).

Numerical experiments were conducted in the case where M is a hyperbolic space of curvature equal to -1 and of dimension d. The following table was obtained for the values $\sigma^2 = \tau^2 = 0.1$, using samples $(x_1, \ldots, x_N)$ where $N = 2 \times 10^5$.

dimension d	2	3	4	5	6	7	8	9	10
$\bar{m}_1(\hat{x}_{\text{MAP}})$	0.28	0.35	0.41	0.47	0.50	0.57	0.60	0.66	0.70
$d(\bar{x}_{\text{MMS}}, \hat{x}_{\text{MAP}})$	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.02	0.03

and the following table for $\sigma^2 = 1$ and $\tau^2 = 0.5$, again using $N = 2 \times 10^5$.

dimension d	2	3	4	5	6	7	8	9	10
$\bar{m}_1(\hat{x}_{\text{MAP}})$	0.75	1.00	1.12	1.44	1.73	1.97	2.15	2.54	2.91
$d(\bar{x}_{\text{MMS}}, \hat{x}_{\text{MAP}})$	0.00	0.00	0.03	0.02	0.02	0.03	0.04	0.03	0.12

The first table confirms Proposition 3.1. The second table, more surprisingly, shows that $\hat{x}_{\text{MMS}}$ and $\hat{x}_{\text{MAP}}$ can be quite close to each other, even when $\rho \neq 1/2$.

In both of these tables, $d(\bar{x}_{\text{MMS}}, \hat{x}_{\text{MAP}})$ is an approximation of $d(\hat{x}_{\text{MMS}}, \hat{x}_{\text{MAP}})$, based on using the empirical barycentre $\bar{x}_{\text{MMS}}$ instead of $\hat{x}_{\text{MMS}}$. The main source of error affecting this approximation is the fact that the samples $(x_1, \ldots, x_N)$ follow from a Metropolis-Hastings algorithm, and not directly from the posterior density π.

Other values of σ^2 and τ^2 lead to similar orders of magnitude for $\bar{m}_1(\hat{x}_{\text{MAP}})$ and $d(\bar{x}_{\text{MMS}}, \hat{x}_{\text{MAP}})$. While $\bar{m}_1(\hat{x}_{\text{MAP}})$ increases with the dimension d, $d(\bar{x}_{\text{MMS}}, \hat{x}_{\text{MAP}})$ does not appear sensitive to increasing dimension.

Based on these experimental results, one is tempted to conjecture that $\hat{x}_{\text{MMS}} = \hat{x}_{\text{MAP}}$, even when $\rho \neq 1/2$. Of course, numerical experiments do not equate to a mathematical proof.

3.3 Computing the MMS

3.3.1 Metropolis-Hastings algorithm

A crucial step in Bayesian inference is sampling from the posterior density. Here, this is $\pi(x)$ given by (41). Since $\pi(x)$ is known only up to normalisation, a suitable sampling method is afforded by the Metropolis-Hastings algorithm. This algorithm generates a Markov chain $(x_n\,; n \geq 1)$, with transition kernel [27]

$$Pf(x) \;=\; \int_M \alpha(x,y) q(x,y) f(y) \mathrm{vol}(dy) + \rho(x) f(x) \tag{47}$$

for any bounded measurable function $f : M \to \mathbb{R}$, where $\alpha(x,y)$ is the probability of accepting a transition from x to dy, and $\rho(x)$ is the probability of staying at x, and where $q(x,y)$ is the proposed transition density

$$q(x,y) \geq 0 \;\text{ and }\; \int_M q(x,y)\mathrm{vol}(dy) = 1 \quad \text{for } x \in M \tag{48}$$

In the following, (x_n) will always be an isotropic Metropolis-Hastings chain, in the sense that $q(x,y) = q(d(x,y))$, so $q(x,y)$ only depends on the distance $d(x,y)$. In this case, the acceptance probability $\alpha(x,y)$ is given by $\alpha(x,y) = \min\{1, \pi(y)/\pi(x)\}$.

The aim of the Metropolis-Hastings algorithm is to produce a Markov chain (x_n) which is geometrically ergodic. Geometric ergodicity means the distribution π_n of x_n converges to π, with a geometric rate, in the sense that there exist $\beta \in (0,1)$ and $R(x_1) \in (0,\infty)$, as well as a function $V : M \to \mathbb{R}$, such that (in the following, $\pi(dx) = \pi(x)\mathrm{vol}(dx)$)

$$V(x) \geq \max\left\{1, d^2(x,x^*)\right\} \text{ for some } x^* \in M \tag{49}$$

$$\left|\int_M f(x)(\pi_n(dx) - \pi(dx))\right| \leq R(x_1)\beta^n \tag{50}$$

for any function $f : M \to \mathbb{R}$ with $|f| \leq V$. If the chain (x_n) is geometrically ergodic, then it satisfies the strong law of large numbers [28]

$$\frac{1}{N}\sum_{n=1}^{N} f(x_n) \longrightarrow \int_M f(x)\pi(dx) \quad \text{(almost-surely)} \tag{51}$$

as well as a corresponding central limit theorem (see Theorem 17.0.1, in [28]). Then, in practice, the Metropolis-Hastings algorithm can be used to generate samples (x_n) from the posterior density $\pi(x)$.

The following general statement can be proved, concerning the geometric ergodicity of isotropic Metropolis-Hastings chains. The proof (see [13], Section 4.6) is a generalisation of the one carried out in the special case where M is a Euclidean space [29].

Proposition 3.3. *Let M be a Riemannian symmetric space, which belongs to the non-compact case. Assume $(x_n\,; n \geq 1)$ is a Markov chain in M, with transition kernel given by (47), with proposed transition density $q(x,y) = q(d(x,y))$, and with strictly positive invariant density π.*

The chain (x_n) satisfies (49) and (50), if the following assumptions hold,
(a1) there exists $x^* \in M$, such that $r(x) = d(x^*,x)$ and $\ell(x) = \log\pi(x)$ satisfy

$$\limsup_{r(x)\to\infty} \frac{\langle \mathrm{grad}\, r, \mathrm{grad}\, \ell\rangle_x}{r(x)} \;<\; 0$$

(a2) if $n(x) = \mathrm{grad}\,\ell(x)/\|\mathrm{grad}\,\ell(x)\|$, then $n(x)$ satisfies

$$\limsup_{r(x)\to\infty} \langle \mathrm{grad}\, r, n\rangle_x \;<\; 0$$

(a3) there exist $\delta_q > 0$ and $\varepsilon_q > 0$ such that $d(x,y) < \delta_q$ implies $q(x,y) > \varepsilon_q$

Remark : the posterior density π in (41) verifies Assumptions (a1) and (a2). To see this, let $x^* = z$, and write

$$\mathrm{grad}\,\ell(x) \;=\; -\frac{1}{\tau^2} r(x)\mathrm{grad}\, r(x) - \frac{1}{\sigma^2}\mathrm{grad}\, f_y(x)$$

where $f_y(x) = d^2(y,x)/2$. Then, taking the scalar product with $\operatorname{grad} r$,

$$\langle \operatorname{grad} r, \operatorname{grad} \ell \rangle_x = -\frac{1}{\tau^2} r(x) - \frac{1}{\sigma^2} \langle \operatorname{grad} r, \operatorname{grad} f_y \rangle_x \tag{52}$$

since $\operatorname{grad} r(x)$ is a unit vector, for all $x \in M$. Now, $\operatorname{grad} f_y(x) = -\operatorname{Exp}_x^{-1}(y)$ (see [30]). But, since $r(x)$ is a convex function of x, it follows, by (85) in Appendix B.1, that

$$\langle \operatorname{grad} r, \operatorname{Exp}_x^{-1}(y) \rangle \;\leq\; r(y) - r(x)$$

for any $y \in M$. Thus, the right-hand side of (52) is strictly negative, as soon as $r(x) > r(y)$, and Assumption (a1) is indeed verified. That Assumption (a2) is also verified can be proved by a similar reasoning. ■

Remark : on the other hand, Assumption (a3) holds, if the proposed transition density $q(x,y)$ is a Gaussian density, $q(x,y) = p(y|x,\tau_q)$. With this choice of $q(x,y)$, all the assumptions of Proposition 3.3 are verified, for the posterior density π in (41). Therefore, Proposition 3.3 implies that the Metropolis-Hastings algorithm generates geometrically ergodic samples $(x_n\,; n \geq 1)$, from this posterior density. ■

3.3.2 The empirical barycentre

Let $(x_n\,; n \geq 1)$ be a Metropolis-Hastings Markov chain in M, with its transition kernel (47), and invariant density π. Assume the chain (x_n) is geometrically ergodic, so it satisfies the strong law of large numbers (51).

Then, let $\bar{x}_N$ denote the empirical barycentre of the first N samples $(x_1, \ldots, x_N)$. This is the unique global minimum of the variance function

$$\mathcal{E}_N(w) \;=\; \frac{1}{2N} \sum_{n=1}^{N} d^2(w, x_n) \tag{53}$$

Let $\hat{x}$ denote the Riemannian barycentre of the invariant density π. It turns out that $\bar{x}_N$ converges almost-surely to $\hat{x}$.

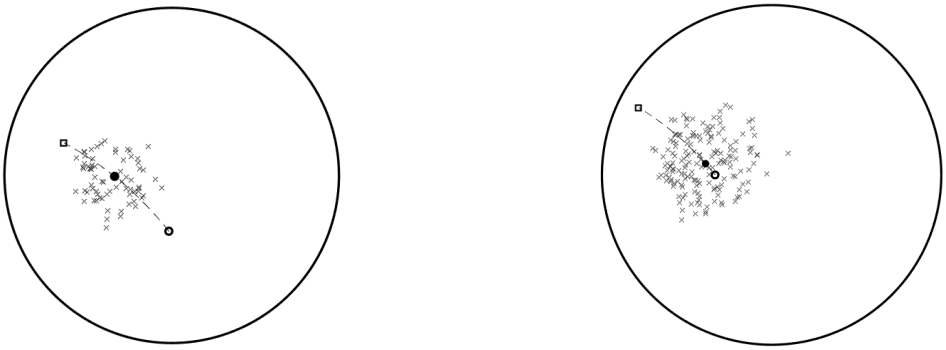
Proposition 3.4. *Let (x_n) be any Markov chain in a Hadamard manifold M, with invariant distribution π. Denote $\bar{x}_N$ the empirical barycentre of $(x_1, \ldots, x_N)$, and $\hat{x}$ the Riemannian barycentre of π. If (x_n) satisfies the strong law of large numbers (51), then $\bar{x}_N$ converges to $\hat{x}$, almost-surely.*

The proof of Proposition 3.4 is nearly a word-for-word repetition of the proof in [31] (that of Theorem 2.3).

According to the remarks after Proposition 3.3, the Metropolis-Hastings Markov chain (x_n), whose invariant density is the posterior density $\pi(x)$, given by (41), is geometrically ergodic. Therefore, by Proposition 3.4, the empirical barycentre $\bar{x}_{\text{MMS}}$, of the samples $(x_1, \ldots, x_N)$, converges almost-surely to the minimum mean square error estimator $\hat{x}_{\text{MMS}}$ (since this is just the barycentre of the posterior density π). This provides a practical strategy for approximating $\hat{x}_{\text{MMS}}$. Indeed, $\bar{x}_{\text{MMS}}$ can be computed using the Riemannian gradient descent method (this method is discussed in Appendix B.4).

Remark : this strategy for approxmating $\hat{x}_{\text{MMS}}$ provided the numerical results discussed

in Paragraph 3.2. For an additional, visual illustration, consider (as in Paragraph 3.2) the case where M is a space of constant negative curvature -1, and of dimension $d = 2$. Figure 1 represents M in the shape of the Poincaré disc. The prior barycentre z is designated by a square $\square$ and the observation y by a circle $\circ$. Grey crosses $\times$ mark the last 1000 out of $N = 100000$ samples x_n generated using the Metropolis-Hastings kernel (47), and the empirical barycentre $\bar{x}_{\text{MMS}}$ is designated by a black circle $\bullet$. In both of the Subfigures 1a and 1b, $\bar{x}_{\text{MMS}}$ is seen to lie on the geodesic connecting z and y, here the dashed circle arc. Note that 1a corresponds to Proposition 3.1 and 1b to Proposition 3.2.



(a) $\sigma^2 = \tau^2 = 0.1$ (Proposition 3.1) (b) $\sigma^2 = 0.1\,; \tau^2 = 1$ (Proposition 3.2)

Figure 1: Poincaré disc with $z = \square\,; y = \circ\,; \bar{x}_{\text{MMS}} = \bullet$

3.4 Proof of Proposition 3.2

Neither this proposition, nor its proof, appeared in [13]. The proof is here given in a series of lemmas. Recall that M is now a hyperbolic space (simply-connected space of constant negative curvature -1).

Lemma 3.1. *Let $\gamma : \mathbb{R} \to M$ denote the geodesic curve with $\gamma(0) = z$ and $\gamma(1) = y$. Then, $\hat{x}_{\text{MMS}}$ lies on this geodesic curve γ.*

Proof: recall from [32] (Section 3) that there exists an isometry $\sigma : M \to M$, such that $\sigma \circ \sigma$ is the identity (σ is an involution), and the set of fixed points of σ is exactly the geodesic curve γ. The key point in the following, which can be seen from (41), is that

$$\pi(\sigma(x)) = \pi(x) \qquad \text{for } x \in M \tag{54}$$

In other words, σ leaves invariant the posterior density π. Let $\mathcal{E}_\pi$ be the function in (43). Then, note that

$$(\mathcal{E}_\pi \circ \sigma)(w) = \frac{1}{2}\int_M d^2(w, \sigma(x))\pi(x)\text{vol}(dx) = \frac{1}{2}\int_M d^2(w, x)\pi(\sigma(x))\sigma^*(\text{vol})(dx)$$

where the first equality follows from (41), because σ is an isometry and an involution, and the second equality by a change of variables ($\sigma^*(\text{vol})$ denotes the pullback of the volume form vol by σ). Using (54) and the fact that σ preserves the volume, it now follows that

$$(\mathcal{E}_\pi \circ \sigma)(w) = \mathcal{E}_\pi(w) \qquad \text{for } w \in M \tag{55}$$

Finally, taking $w = \hat{x}_{\text{MMS}}$ and recalling that $\hat{x}_{\text{MMS}}$ is the unique global minimiser of $\mathcal{E}_\pi$, it follows that $\sigma(\hat{x}_{\text{MMS}}) = \hat{x}_{\text{MMS}}$, so that $\hat{x}_{\text{MMS}}$ indeed lies on γ. ■

Lemma 3.2. *There exists a continuous function $c : \mathbb{R} \to \mathbb{R}$ such that*

$$\operatorname{grad} \mathcal{E}_\pi(\gamma(t)) = c(t)\dot{\gamma}(t) \qquad \textit{for } t \in \mathbb{R} \tag{56}$$

Remark : taking covariant derivatives in (56),

$$\operatorname{Hess} \mathcal{E}_\pi(\gamma(t)) \cdot \dot{\gamma}(t) = c'(t)\dot{\gamma}(t) \tag{57}$$

where $c'(t) = dc(t)/dt$. Because $\mathcal{E}_\pi$ is 1-strongly convex (see Item (iii) of Proposition B.2, Appendix B.1), it follows that $c'(t) \geq 1$. In particular, $c(t)$ is strictly increasing.

Proof : let $w = \gamma(t)$ and take the gradient of (55). This yields

$$d\sigma \cdot \operatorname{grad} \mathcal{E}_\pi(\gamma(t)) = \operatorname{grad} \mathcal{E}_\pi(\gamma(t)) \tag{58}$$

However, the derivative $d\sigma : T_{\gamma(t)}M \to T_{\gamma(t)}M$ is equal to 1 on vectors parallel to $\dot{\gamma}(t)$ and to -1 on vectors orthogonal to $\dot{\gamma}(t)$. Thus, (58) implies that $\operatorname{grad} \mathcal{E}_\pi(\gamma(t))$ should be parallel to $\dot{\gamma}(t)$. This is equivalent to (56). ■

The next step in the proof will be to compute $c(0)$ and $c(1)$. This will show that $c(0)$ is negative and $c(1)$ is positive. Computing $c(0)$ and $c(1)$ requires taking a closer look at the posterior density π.

Lemma 3.3. *Let $Z(z, y, \tau, \sigma)$ denote the missing normalising factor in (41). Then, $Z(z, y, \tau, \sigma) = Z(\delta, \tau, \sigma)$, where $\delta = d^2(z, y)$.*

Proof : a hyperbolic space is a two-point homogeneous space [15] (Page 355). This means if $z', y' \in M$ have $d(z', y') = d(z, y)$, then there exists an isometry $g : M \to M$ such $g(z) = z'$ and $g(y) = y'$. Now, since g is an isometry,

$$\begin{aligned} Z(z, y, \tau, \sigma) &= \int_M \exp\left[-\frac{d^2(y, x)}{2\sigma^2} - \frac{d^2(x, z)}{2\tau^2}\right] \operatorname{vol}(dx) \\ &= \int_M \exp\left[-\frac{d^2(g(y), g(x))}{2\sigma^2} - \frac{d^2(g(x), g(z))}{2\tau^2}\right] \operatorname{vol}(dx) \end{aligned}$$

Thus, introducing the change of variables $w = g(x)$,

$$Z(z, y, \tau, \sigma) = \int_M \exp\left[-\frac{d^2(y', w)}{2\sigma^2} - \frac{d^2(w, z')}{2\tau^2}\right] \operatorname{vol}(dw) = Z(z', y', \tau, \sigma)$$

In other words, $Z(z, y, \tau, \sigma)$ only depends on the distance between z and y. ■

It is now possible to compute $c(0)$ and $c(1)$.

Lemma 3.4. *There exists a positive constant $\psi(\delta, \tau, \sigma)$, such that*

$$c(0) = -\psi(\delta, \tau, \sigma)\tau^2 \;\; \textit{and} \;\; c(1) = \psi(\delta, \tau, \sigma)\sigma^2 \tag{59}$$

In the notation of (56).

Proof: for any value of the parameters (z, y, τ, σ),

$$\int_M \exp\left[-\frac{d^2(y,x)}{2\sigma^2} - \frac{d^2(x,z)}{2\tau^2} - \log Z(\delta,\tau,\sigma)\right] \mathrm{vol}(dx) = 1$$

where $\delta = d^2(z,y)$. Taking the gradient of this identity with respect to z, and using

$$\mathrm{grad}_z d^2(x,z) = -2\mathrm{Exp}_z^{-1}(x) \quad ; \quad \mathrm{grad}_z d^2(z,y) = -2\mathrm{Exp}_z^{-1}(y)$$

where grad_z denotes the gradient with respect to z, it follows that

$$\frac{1}{\tau^2}\int_M \mathrm{Exp}_z^{-1}(x)\pi(x)\mathrm{vol}(dx) - \psi(\delta,\tau,\sigma)\,\mathrm{Exp}_z^{-1}(y) = 0 \tag{60}$$

where $\psi(\delta,\tau,\sigma) = -2 \times \partial \log Z(\delta,\tau,\sigma)/\partial\delta$. However, here one has,

$$\int_M \mathrm{Exp}_z^{-1}(x)\pi(x)\mathrm{vol}(dx) = -\mathrm{grad}\,\mathcal{E}_\pi(z) \quad ; \quad \mathrm{Exp}_z^{-1}(y) = \dot\gamma(0) \tag{61}$$

Thus, replacing (61) into (60), it follows that

$$c(0) = -\psi(\delta,\tau,\sigma)\tau^2$$

which is the first part of (59). The second part can be proved in the same way, taking the gradient with respect to y rather than z. The fact that $\psi(\delta,\tau,\sigma) > 0$ follows because $c(t)$ is strictly increasing (see the remark after Lemma 3.2), and has at most one zero (because $\mathcal{E}_\pi$ has exactly one stationary point). ■

It is now possible to complete the proof of Proposition 3.2. Lemma 3.4, shows that $c(0)$ is negative and $c(1)$ is positive. Therefore, $c(t^*) = 0$ for some $t^* \in (0,1)$. By (56) of Lemma 3.2, $\mathrm{grad}\,\mathcal{E}_\pi(\gamma(t^*)) = 0$. Since $\mathcal{E}_\pi$ is strongly convex, it follows immediately that $\gamma(t^*)$ is the unique global minimiser of $\mathcal{E}_\pi$. In other words, $\hat{x}_{\mathrm{MMS}} = \gamma(t^*)$, as required. ■

A Riemannian symmetric spaces

A Riemannian symmetric space is a Riemannian manifold M, such that, for each $x \in M$, there exists an isometry $s_x : M \to M$, with $s_x(x) = x$ and $ds_x(x) = -\mathrm{Id}_x$. This isometry s_x is called the geodesic symmetry at x [15].

Let G denote the identity component of the isometry goup of M, and $K = K_o$ be the stabiliser in G of some point $o \in M$. Then, $M = G/K$ is a Riemannian homogeneous space. The mapping $\theta : G \to G$, $\theta(g) = s_o \circ g \circ s_o$ is an involutive isomorphism of G.

Let $\mathfrak{g}$ denote the Lie algebra of G, and consider the Cartan decomposition, $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$, where $\mathfrak{k}$ is the $+1$ eigenspace of $d\theta$ and $\mathfrak{p}$ is the -1 eigenspace of $d\theta$. One clearly has the commutation relations,

$$[\mathfrak{k}, \mathfrak{k}] \subset \mathfrak{k} \ ; \ [\mathfrak{k}, \mathfrak{p}] \subset \mathfrak{p} \ ; \ [\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{k} \tag{62}$$

In addition, it turns out that $\mathfrak{k}$ is the Lie algebra of K, and that $\mathfrak{p}$ may be identified with T_oM, in a natural way.

The Riemannian metric of M may always be expressed in terms of an $\mathrm{Ad}(K)$-invariant scalar product Q on $\mathfrak{g}$. If $x \in M$ is given by $x = g \cdot o$ for some $g \in G$ (where $g \cdot o = g(o)$), then

$$\langle u,v \rangle_x = Q(g^{-1} \cdot u, g^{-1} \cdot v) \tag{63}$$

where the vectors $g^{-1} \cdot u$ and $g^{-1} \cdot v$, which belong to T_oM, are identified with elements of $\mathfrak{p}$. Here, by an abuse of notation, $dg^{-1} \cdot u$ is denoted $g^{-1} \cdot u$.

Let $\exp : \mathfrak{g} \to G$ denote the Lie group exponential. If $v \in T_oM$, then the Riemannian exponential $\mathrm{Exp}_o(v)$ is given by

$$\mathrm{Exp}_o(v) = \exp(v) \cdot o \tag{64}$$

Moreover, if Π_0^t denotes parallel transport along the geodesic $c(t) = \mathrm{Exp}_o(tv)$, then

$$\Pi_0^t(u) = \exp(tv) \cdot u \tag{65}$$

for any $u \in T_oM$ (note that the identification $T_oM \simeq \mathfrak{p}$ is always made, implicitly). Using (65), one can derive the following expression for the Riemann curvature tensor at o,

$$R_o(v,u)w = -[[v,u],w] \qquad v,u,w \in T_oM \tag{66}$$

A fundamental property of symmetric spaces is that the curvature tensor is parallel : $\nabla R = 0$. This is often used to solve the Jacobi equation [15][14], and then express the derivative of the Riemannian exponential,

$$\mathrm{dExp}_x(v)(u) \ = \ \exp(v) \cdot \mathrm{sh}(R_v)(u) \tag{67}$$

where $\mathrm{sh}(R_v) = \sum_{n=0}^{\infty} (R_v)^n/(2n+1)!$ for the self-adjoint curvature operator $R_v(u) = [v,[v,u]]$. As a result of (67), since $\exp(v)$ is an isometry, the following expression of the Riemannian volume is immediate

$$\mathrm{Exp}_o^*(\mathrm{vol}) = |\det(\mathrm{sh}(R_v))| \, dv \tag{68}$$

where dv denotes the volume form on T_oM, associated with the restriction of the scalar product Q to $\mathfrak{p}$.

Expression (68) yields applicable integral formulae, when $\mathfrak{g}$ is a reductive Lie algebra ($\mathfrak{g} = \mathfrak{z} + \mathfrak{g}_{ss}$: $\mathfrak{z}$ the centre of $\mathfrak{g}$ and $\mathfrak{g}_{ss}$ semisimple). If $\mathfrak{a}$ is a maximal Abelian subspace of $\mathfrak{p}$, any $v \in \mathfrak{p}$ is of the form $v = \mathrm{Ad}(k)\, a$ for some $k \in K$ and $a \in \mathfrak{a}$ (see [15], Lemma 6.3, Chapter V). Moreover, using the fact that $\mathrm{Ad}(k)$ is an isomorphism of $\mathfrak{g}$,

$$\mathrm{Ad}(k^{-1}) \circ R_v \circ \mathrm{Ad}(k) = R_a = \sum_{\lambda \in \Delta_+} (\lambda(a))^2 \, \Pi_\lambda \tag{69}$$

where each $\lambda \in \Delta_+$ is a linear form $\lambda : \mathfrak{a} \to \mathbb{R}$, and Π_λ is the orthogonal projector onto the corresponding eigenspace of R_a. Here, Δ_+ is the set of positive roots of $\mathfrak{g}$ with respect to $\mathfrak{a}$ [15] (see Lemma 2.9, Chapter VII).

It is possible to use the diagonalisation (69), in order to evaluate the determinant (68). To obtain a regular parameterisation, let $S = K/K_{\mathfrak{a}}$, where $K_{\mathfrak{a}}$ is the centraliser of $\mathfrak{a}$ in K. Then, let $\varphi : S \times \mathfrak{a} \to M$ be given by $\varphi(s, a) = \mathrm{Exp}_o(\beta(s, a))$ where $\beta(s, a) = \mathrm{Ad}(s)\, a$. Now, by (68) and (69),

$$\varphi^*(\mathrm{vol}) = \prod_{\lambda \in \Delta_+} \left| \frac{\sinh \lambda(a)}{\lambda(a)} \right|^{m_\lambda} \beta^*(dv)$$

where m_λ is the multiplicity of λ (the rank of Π_λ). On the other hand, one may show

$$\beta^*(dv) = \prod_{\lambda \in \Delta_+} |\lambda(a)|^{m_\lambda} \, da \, \omega(ds) \tag{70}$$

where da is the volume form on $\mathfrak{a}$, and ω is the invariant volume induced onto S from K.

Finally, the Riemannian volume, in terms of the parameterisation φ, can be expressed in the following way

$$\varphi^*(\mathrm{vol}) = \prod_{\lambda \in \Delta_+} |\sinh \lambda(a)|^{m_\lambda} \, da \, \omega(ds) \tag{71}$$

Using (71), it will be possible to write down integral formulae for Riemannian symmetric spaces, either non-compact or compact.

A.1 The non-compact case

This is the case were $\mathfrak{g}$ admits an $\mathrm{Ad}(G)$-invariant, non-degenerate, symmetric bilinear form B, such that $Q(u, z) = -B(u, d\theta(z))$ is an $\mathrm{Ad}(K)$-invariant scalar product on $\mathfrak{g}$.

In this case, B is negative-definite on $\mathfrak{k}$ and positive-definite on $\mathfrak{p}$. Moreover, the linear map $\mathrm{ad}(z) = [z, \cdot]$ is skew-symmmetric or symmetric (with respect to Q), according to whether $z \in \mathfrak{k}$ or $z \in \mathfrak{p}$.

If $u_1, u_2 \in \mathfrak{p}$ are orthonormal, the sectional curvature of $\mathrm{Span}(u_1, u_2)$ is found from (66), $\kappa(u_1, u_2) = -\|[u_1, u_2]\|_o^2 \leq 0$. Therefore, M has non-positive sectional curvatures.

In fact, M is a Hadamard manifold. It is geodesically complete by (64). It is moreover simply connected, because $\mathrm{Exp}_o : \mathfrak{p} \to M$ is a diffeomorphism [15] (Theorem 1.1, Chapter VI). Thus, (68) yields a first integral formula,

$$\int_M f(x) \, \mathrm{vol}(dx) = \int_{\mathfrak{p}} f(\mathrm{Exp}_o(v)) \, |\det(\mathrm{sh}(R_v))| \, dv \tag{72}$$

To obtain an integral formula from (71), one should first note that $\beta : S \times \mathfrak{a} \to \mathfrak{p}$ is not regular, nor one-to-one. Recall the following:
• the hyperplanes $\lambda(a) = 0$, where $\lambda \in \Delta_+$, divide $\mathfrak{a}$ into finitely many connected components, which are open and convex sets, known as Weyl chambers. From (70), β is regular on each Weyl chamber.
• let $K'_{\mathfrak{a}}$ denote the normaliser of $\mathfrak{a}$ in K. Then, $W = K'_{\mathfrak{a}}/K_{\mathfrak{a}}$ is a finite group of automorphisms of $\mathfrak{a}$, called the Weyl group, which acts freely transitively on the set of Weyl chambers [15] (Theorem 2.12, Chapter VII).

Then, for each Weyl chamber C, β is regular and one-to-one, from $S \times C$ onto its image in $\mathfrak{p}$. Moreover, if $\mathfrak{a}_r$ is the union of the Weyl chambers ($a \in \mathfrak{a}_r$ if and only if $\lambda(a) \neq 0$ for any $\lambda \in \Delta_+$), then β is regular and $|W|$-to-one from $S \times \mathfrak{a}_r$ onto its image in $\mathfrak{p}$. To obtain the desired integral formula, it only remains to note that φ is a diffeomorphism from $S \times C$ onto its image in M. However, this image is the set M_r of regular values of φ. By Sard's lemma, its complement is negligible [33].

Proposition A.1. *Let $M = G/K$ be a Riemannian symmetric space, which belongs to the "non-compact case", just described. Then, for any bounded continuous function $f : M \to \mathbb{R}$,*

$$\int_M f(x)\,\mathrm{vol}(dx) = \int_{C_+}\int_S f(\varphi(s,a)) \prod_{\lambda\in\Delta_+} (\sinh \lambda(a))^{m_\lambda}\, da\,\omega(ds) \tag{73}$$

$$= \frac{1}{|W|}\int_{\mathfrak{a}}\int_S f(\varphi(s,a)) \prod_{\lambda\in\Delta_+} |\sinh \lambda(a)|^{m_\lambda}\, da\,\omega(ds) \tag{74}$$

Here, C_+ is the Weyl chamber $C_+ = \{a \in \mathfrak{a} : \lambda \in \Delta_+ \Rightarrow \lambda(a) > 0\}$.

A.2 The compact case

In this case, $\mathfrak{g}$ admits an $\mathrm{Ad}(G)$-invariant scalar product Q. Therefore, $\mathrm{ad}(z)$ is skew-symmmetric, with respect to Q, for each $z \in \mathfrak{g}$. Using (66), it follows that M is compact, with non-negative sectional curvature.

In fact, the compact case may be obtained from the previous non-compact case by *duality*. Denote $\mathfrak{g}_{\mathbb{C}}$ the complexification of $\mathfrak{g}$, and let $\mathfrak{g}^* = \mathfrak{k} + \mathfrak{p}_*$ where $\mathfrak{p}_* = i\mathfrak{p}$. Then, $\mathfrak{g}^*$ is a compact real form of $\mathfrak{g}_{\mathbb{C}}$ (that is, $\mathfrak{g}^*$ is a compact Lie algebra, and its complexification is equal to $\mathfrak{g}_{\mathbb{C}}$). Denote G^* the connected Lie group with Lie algebra $\mathfrak{g}^*$.

If $M = G/K$ is a Riemannian symmetric space which belongs to the non-compact case, then $M^* = G^*/K$ is a Riemannian symmetric space which belongs to the compact case. Formally, to pass from the non-compact case to the compact case, all one has to do is replace a by ia. Applying this recipe to (71), one obtains

$$\varphi^*(\mathrm{vol}) = \prod_{\lambda\in\Delta_+} |\sin \lambda(a)|^{m_\lambda}\, da\,\omega(ds) \tag{75}$$

where da is the volume form on $\mathfrak{a}_* = i\mathfrak{a}$, and ω is the invariant volume induced onto S from K. Note that the image under Exp_o of $\mathfrak{a}_*$ is the torus $T_* = \mathfrak{a}_*/\mathfrak{a}_K$, where $\mathfrak{a}_K$ is the lattice given by $\mathfrak{a}_K = \{a \in \mathfrak{a}_* : \mathrm{Exp}_o(a) = o\}$. Recall the following:

• $\varphi(s,a)$ only depends on $t = \mathrm{Exp}_o(a)$. Thus, φ may be considered as a map from $S \times T_*$ to M.
• if $a \in \mathfrak{a}_K$ then $\exp(2a) = e$ (the identity element in G^*). Thus, $\lambda(a) \in i\pi\,\mathbb{Z}$ for all $\lambda \in \Delta_+$ [15] (Page 383). Therefore, there exists a function $D : T \to \mathbb{R}$, such that

$$D(t) = \prod_{\lambda\in\Delta_+} |\sin\lambda(a)|^{m_\lambda} \qquad \text{whenever } t = \mathrm{Exp}_o(a)$$

Now, T_* is a totally flat submanifold of M. Therefore, $\mathrm{Exp}^*(dt) = da$, where dt denotes the invariant volume induced onto T_* from M. With a slight abuse of notation, (75) now reads,

$$\varphi^*(\mathrm{vol}) = D(t)\,dt\,\omega(ds) \tag{76}$$

Denote $(T_*)_r$ the set of $t \in T_*$ such that $D(t) \neq 0$. By the same arguments as in the non-compact case, φ is a regular $|W|$-to-one map from $S \times (T_*)_r$ onto M_r, the set of regular values of φ.

Proposition A.2. *Let $M = G^*/K$ be a Riemannian symmetric space, which belongs to the "compact case", just described. For any bounded continuous function $f : M \to \mathbb{R}$,*

$$\int_M f(x)\,\mathrm{vol}(dx) \;=\; \frac{1}{|W|}\int_{T_*}\int_S f(\varphi(t,a))\,D(t)\,dt\,\omega(ds) \tag{77}$$

A.3 Example of Propositions A.1 and A.2

consider $M = \mathrm{H}(N)$ the space of $N \times N$ Hermitian positive-definite matrices. Here, $G = \mathrm{GL}(N,\mathbb{C})$ and $K = U(N)$. Moreover, $B(u,z) = \mathrm{Re}(\mathrm{tr}(uz))$ and $d\theta(z) = -z^\dagger$. Thus, $\mathfrak{p}$ is the space of $N \times N$ Hermitian matrices, and one may choose $\mathfrak{a}$ the space of $N \times N$ real diagonal matrices. The positive roots are the linear maps $\lambda(a) = a_{ii} - a_{jj}$ where $i < j$, and each one has its multiplicity equal to 2. The Weyl group W is the group of permutation matrices in $U(N)$ (so $|W| = N!$). Finally, $S = U(N)/T_N \equiv S_N$, where T_N is the torus of diagonal unitary matrices. By (74),

$$\int_{\mathrm{H}(N)} f(x)\,\mathrm{vol}(dx) \;=\; \frac{1}{N!}\int_{\mathfrak{a}}\int_{S_N} f\left(s\exp(2a)s^\dagger\right)\prod_{i<j}\sinh^2(a_{ii}-a_{jj})\,da\,\omega(ds) \tag{78}$$

where $da = da_{11}\ldots da_{NN}$. Now, assume f is a class function: $f(k\cdot x) = f(x)$ for $k \in K$ and $x \in \mathrm{H}(N)$. That is, $f(x)$ depends only on the eigenvalues $x_i = e^{r_i}$ of x. By (78),

$$\int_{\mathrm{H}(N)} f(x)\,\mathrm{vol}(dx) \;=\; \frac{\omega(S_N)}{2^N N!}\int_{\mathbb{R}^N} f\left(\exp(r)\right)\prod_{i<j}\sinh^2((r_i-r_j)/2)\,dr \tag{79}$$

or, by introducing the eigenvalues x_i as integration variables,

$$\int_{\mathrm{H}(N)} f(x)\,\mathrm{vol}(dx) \;=\; \frac{\omega(S_N)}{2^{N^2} N!}\int_{\mathbb{R}_+^N} f\left(x_1,\ldots,x_N\right)|V(x)|^2\prod_{i=1}^N x_i^{-N}\,dx_i \tag{80}$$

where $V(x) = \prod_{i<j}(x_j - x_i)$ is the Vandermonde determinant.

The dual of H(N) is the unitary group $U(N)$. Here, $G^* = U(N) \times U(N)$ and $K \simeq U(N)$, is the diagonal group $K = \{(x,x)\,;x \in U(N)\}$. The Riemannian metric is given by the trace scalar product $Q(u,z) = -\mathrm{tr}(uz)$. Moreover, $T_* = T_N$ and $S = S_N$ (this is $U(N)/T_N$). The positive roots are $\lambda(ia) = a_{ii} - a_{jj}$ where $i < j$ and where a is $N \times N$, real and diagonal[4]. By writing the integral over T_N as a multiple integral, (77) reads,

$$\int_{U(N)} f(x)\,\mathrm{vol}(dx) \;=\; \frac{1}{N!}\int_{[0,2\pi]^N}\int_{S_N} f\left(s\exp(2ia)s^\dagger\right)\prod_{i<j}\sin^2(a_{ii}-a_{jj})\omega(ds)\,da \quad (81)$$

where $da = da_{11}\ldots da_{NN}$.

Now, assume f is a class function. That is, $f(x)$ depends only on eigenvalues $e^{i\theta_i}$ of x. Integrating out s, from (81), it follows,

$$\int_{U(N)} f(x)\,\mathrm{vol}(dx) \;=\; \frac{\omega(S_N)}{2^N N!}\int_{[0,2\pi]^N} f\left(\exp(i\theta)\right)\prod_{i<j}\sin^2((\theta_i-\theta_j)/2)\,d\theta \qquad (82)$$

or, after an elementary manipulation,

$$\int_{U(N)} f(x)\,\mathrm{vol}(dx) \;=\; \frac{\omega(S_N)}{2^{N^2} N!}\int_{[0,2\pi]^N} f\left(\theta_1,\ldots,\theta_N\right)|V(e^{i\theta})|^2\,d\theta_1\ldots\theta_N \qquad (83)$$

where $V(e^{i\theta}) = \prod_{i<j}(e^{i\theta_j} - e^{i\theta_i})$ is the Vandermonde determinant.

Integrals such as (80) and (83) are familiar in random matrix theory [17][34]. The resemblance between these integrals (for example, in the rôle played by the Vandermonde determinant) is at the origin of the sort of "duality" described in Paragraph 2.8.

[4]Please do no confuse the imaginary number i with the subscript i.

B Convex optimisation

B.1 Convex sets and functions

In Euclidean geometry, a convex set A is any set which satisfies the definition : if points x and y belong to A, then the straight line segment between x and y lies entirely in A. One hopes to extend this definition to Riemannian geometry, by letting geodesics play the rôle of straight lines. However, this does not lead to one, but to multiple definitions of a convex set. The present article will focus on the following [30].

Definition B.1. *A subset A of a complete Riemannian manifold M is called strongly convex if, whenever points x and y belong to A, there exists a unique length-minimising geodesic $\gamma_{x,y}$ connecting x and y, and $\gamma_{x,y}$ lies entirely in A.*

Remark : as an example of a different way of defining a convex set, consider the following. A subset A of M is called weakly convex if, whenever points x and y belong to A, there exists a unique geodesic γ in M, such that γ connects x and y, and γ lies entirely in A (this coincides with the definition in [30], because γ is then the unique length-minimising curve, among all curves that connect x and y and lie entirely in A). ■

In Euclidean geometry, a ball of any radius is convex. In a Riemannian manifold, a ball may fail to be strongly (or even weakly) convex, if its radius is too large. On the other hand, a ball with sufficiently small radius is always strongly convex [30][14].

Proposition B.1. *Assume the sectional curvatures of M are bounded above by $\kappa_{\max}^2 \geq 0$. Then, denoting* $\mathrm{inj}(M)$ *the injectivity radius of M, let*

$$R_c(M) = \min\left\{\frac{1}{2}\mathrm{inj}(M), \frac{\pi}{2\kappa_{\max}}\right\} \tag{84}$$

For any $x \in M$ and $R < R_c(M)$, the open ball $B(x, R)$ is strongly convex (if $\kappa_{\max} = 0$, it should be understood that division by zero yields infinity).

Remark : if $\frac{1}{2}\mathrm{inj}(M)$ is replaced by $\mathrm{inj}(M)$ in (84), then $R \leq R_c(M)$ implies $B(x, R)$ is weakly convex [30][35]. ■

There is a certain class of Riemannian manifolds, where balls of any radius are strongly convex. Namely, these are Hadamard manifolds. Recall that a Hadamard manifold is a simply connected, complete Riemannian manifold with non-positive sectional curvatures. In particular [36], this implies $\mathrm{inj}(M) = \infty$ and $\kappa_{\max} = 0$, so that $R_c(M) = \infty$.

The following definition of a convex function on a Riemannian manifold directly extends the usual, well-known definition of a convex function on a Euclidean space.

Definition B.2. *Let A be a strongly convex subset of a complete Riemannian manifold M, and $f : A \to \mathbb{R}$. Then, f is called convex (respectively, strictly convex) if $f(\gamma_{x,y}(t))$ is a convex (respectively, strictly convex) function of the time parameter t, for all $x, y \in A$. Further, if there exists $\alpha > 0$ such that $f(\gamma_{x,y}(t))$ is an α-strongly convex function of t, for all $x, y \in A$, then f is called α-strongly convex.*

For differentiable functions, it is possible to write down first-order and second-order characterisations of convexity [26]. Recall that $\dot{\gamma}_{x,y}(0) = \mathrm{Exp}_x^{-1}(y)$, for $x, y \in A$, where

the dot denotes the time derivative and Exp the Riemannian exponential map [36]. In addition, let grad$f$ and Hess$f$ denote the gradient and Hessian of a function f on M (defined with respect to the Riemannian metric and Levi-Civita connection of M).

Proposition B.2. *Let A be a strongly convex subset of a complete Riemannian manifold M, and $f : A \to \mathbb{R}$.*
(i) assume f is differentiable. Then, f is convex, if and only if

$$f(y) - f(x) \geq \langle \mathrm{grad} f(x), \mathrm{Exp}_x^{-1}(y) \rangle_x \qquad \text{for all } x, y \in A \tag{85}$$

Moreover, f is strictly convex if and only if the above inequality is strict whenever $y \neq x$.
(ii) assume f is differentiable. Then f is α-strongly convex, if and only if,

$$f(y) - f(x) \geq \langle \mathrm{grad} f(x), \mathrm{Exp}_x^{-1}(y) \rangle_x + (\alpha/2) d^2(y, x) \qquad \text{for all } x, y \in A \tag{86}$$

(iii) assume f is twice differentiable. Then f is convex if and only if $\mathrm{Hess} f(x) \succeq 0$*, and strictly convex if and only if* $\mathrm{Hess} f(x) \succ 0$*, for all $x \in A$. Moreover, f is α-strongly convex if and only if* $\mathrm{Hess} f(x) \succeq \alpha g(x)$*, for all $x \in A$.*

Here, $\langle \cdot, \cdot \rangle$ and $d(\cdot, \cdot)$ denote the Riemannian scalar product and distance, associated with the Riemannian metric tensor g of M. Moreover, $\succ$ stands for the Loewner order. A straightforward consequence of (ii) in Proposition B.2 is the so-called PL inequality (PL stands for Polyak-Lojasiewicz [37]). This inequality will be used in Paragraph B.4.2.

Proposition B.3. *Let $f : A \to \mathbb{R}$ be a twice differentiable, α-strongly convex function. If f has its minimum at $x^* \in A$, then*

$$\|\mathrm{grad} f(x)\|_x^2 \geq 2\alpha (f(x) - f(x^*)) \qquad \text{for all } x \in A \tag{87}$$

B.2 Second-order Taylor formula

Consider the second-order Taylor formula, for a twice-differentiable function $f : M \to \mathbb{R}$ (as usual, M is a complete Riemannian manifold). For $x \in M$ and $v \in T_x M$,

$$f(\mathrm{Exp}_x(v)) = f(x) + \langle \mathrm{grad} f(x), v \rangle_x + \frac{1}{2} \mathrm{Hess} f_{\gamma(t^*)}(\dot{\gamma}, \dot{\gamma}) \tag{88}$$

where γ is the geodesic curve $\gamma(t) = \mathrm{Exp}_x(tv)$ and $t^* \in (0, 1)$. Formula (88) is the first-order Taylor expansion, with Lagrange remainder, of the function $f(\gamma(t))$, at $t = 0$ [13]. This formula will be the starting point for the study of Riemannian gradient descent in Paragraph B.4. There, it will be applied with $v = -\mu \mathrm{grad} f(x)$ and $\mu \in (0, 1]$.

To apply (88), it is quite helpful to control the second-order term in its right-hand side. One says that f is L-smooth on $B \subset M$, if there exists $L \geq 0$ with $|\mathrm{Hess} f_y(u, u)| \leq L \|u\|_y^2$ for all $y \in B$ and $u \in T_y M$. Then, if $\gamma(t) = \mathrm{Exp}_x(tv)$ belongs to B for all $t \in (0, 1)$,

$$f(\mathrm{Exp}_x(v)) \leq f(x) + \langle \mathrm{grad} f(x), v \rangle_x + (L/2) \|v\|_x^2 \tag{89}$$

Inequality (89) yields the following Proposition B.4. To state this proposition, consider a C^2 function $f : M \to \mathbb{R}$, and assume the sublevel set $B_c = \{x : f(x) \leq c\}$ is compact (and not empty), for some real c. Of course, B_c is contained in some closed ball $B = \bar{B}(z, R)$. Let G be the maximum of $\|\mathrm{grad} f(x)\|_x$, taken over $x \in B$, and $B' = \bar{B}(z, R + G)$. Now, by compactness of B', f is L_c-smooth on B', for some $L_c \geq 0$.

Proposition B.4. *Let $f : M \to \mathbb{R}$ be a C^2 function, with B_c and L_c defined as above, and let $y = \mathrm{Exp}_x(-\mu \mathrm{grad} f(x))$ for some $\mu \in (0, 1]$. If $\mu \leq 1/L_c$, then*

$$f(y) \leq f(x) - (\mu/2)\|\mathrm{grad} f(x)\|_x^2 \qquad \text{for all } x \in B_c \tag{90}$$

In particlar, $x \in B_c$ implies $y \in B_c$.

Remark : as a consequence of (90), if $x^* \in B_c$ is such that $f(x^*)$ is the minimum of $f(x)$, taken over $x \in B_c$, then

$$2L_c(f(x) - f(x^*)) \geq \|\mathrm{grad} f(x)\|_x^2 \qquad \text{for all } x \in B_c \tag{91}$$

which is complementary to (87). ■

B.3 Taylor with retractions

It is customary, in practical applications, to approximate the Riemannian exponential map by another so-called retraction map, which is easier to compute [38]. In this context, it is helpful to derive new versions of Formula (88) and of Proposition B.4, which apply when the exponential map Exp is replaced with a retraction Ret.

Recall that a retraction is a smooth map $\mathrm{Ret} : TM \to M$ (denoted $\mathrm{Ret}(x, v) = \mathrm{Ret}_x(v)$ for $x \in M$ and $v \in T_xM$), such that

$$\mathrm{Ret}_x(0_x) = x \quad \text{and} \quad d\mathrm{Ret}_x(0_x) = \mathrm{Id}_x \tag{92}$$

for all $x \in M$. Here, 0_x is the zero element in T_xM and Id_x is the identity map of T_xM. Most retractions, encountered in practical applications, are regular retractions, in the following sense [13].

Definition B.3. *A retraction $\mathrm{Ret} : TM \to M$ is regular, if there exists a smooth bundle map $\Phi : TM \to TM$, such that*

$$\mathrm{Ret}_x(v) = \mathrm{Exp}_x\left(\Phi_x(v)\right) \qquad \text{for all } x \in M \text{ and } v \in T_xM \tag{93}$$

Here, Φ is denoted $\Phi(x, v) = \Phi_x(v)$ ("bundle map" means $\Phi_x(v) \in T_xM$ for all $v \in T_xM$).

If $\mathrm{Ret} : TM \to M$ is a regular retraction and $f : M \to \mathbb{R}$ is a twice-differentiable function, then (88) and (93) directly imply

$$f\left(\mathrm{Ret}_x(v)\right) = f(x) + \langle \mathrm{grad} f, \Phi_x(v) \rangle_x + \frac{1}{2}\mathrm{Hess} f_{\gamma(t^*)}(\dot\gamma, \dot\gamma) \tag{94}$$

where γ is the geodesic curve $\gamma(t) = \mathrm{Exp}_x(t\Phi_x(v))$ and $t^* \in (0, 1)$. Formula (94) is the required new version of (88).

The Riemannian exponential Exp is a regular retraction, with $\Phi_x = \mathrm{Id}_x$ for $x \in M$. For a general regular retraction Ret, each map $\Phi_x : T_xM \to T_xM$ still agrees with Id_x up to second-order terms [13].

Proposition B.5. *Let $\mathrm{Ret} : TM \to M$ be a regular retraction, with $\Phi : TM \to TM$ given by (93). Then, for each $x \in M$, the map $\Phi_x : T_xM \to T_xM$ verifies*

(a) *$\Phi_x(0_x) = 0_x$ and $\Phi_x'(0_x) = \mathrm{Id}_x$ (the prime denotes the Fréchet derivative).*

(b) *$\Phi_x''(0_x)(v, v) = \ddot c(0)$, where the curve $c(t)$ is given by $c(t) = \mathrm{Ret}_x(tv)$.*

The retraction Ret is called geodesic when $\Phi''_x(0_x)(v,v) = 0$ for $x \in M$ and $v \in T_xM$. In this case, Φ_x agrees with Id_x up to third order terms. For geodesic regular retractions, the following Proposition B.6 provides a new, general version of Proposition B.4.

Proposition B.6. *Let $f : M \to \mathbb{R}$ be a C^2 function, with $B_c = \{x : f(x) \leq c\}$ compact (and not empty), and let* Ret $: TM \to M$ *be a geodesic regular retraction. There exist constants $\beta_c, \delta_c, H_c \geq 0$, which depend on f and* Ret*, such that, for all $x \in B_c$,*

$$f(y) \leq f(x) - \mu\left(1 - (\beta_c H_c/2)\mu - (\delta_c \|\mathrm{grad} f(x)\|_x^2)\mu^2\right)\|\mathrm{grad} f(x)\|_x^2 \tag{95}$$

whenever $y = \mathrm{Ret}_x(-\mu \mathrm{grad} f(x))$ for some $\mu \in (0,1]$. In particular,

$$\frac{1}{2} - (\beta_c H_c/2)\mu - (\delta_c \|\mathrm{grad} f(x)\|_x^2)\mu^2 \geq 0 \Longrightarrow f(y) \leq f(x) - (\mu/2)\|\mathrm{grad} f(x)\|_x^2 \tag{96}$$

Therefore, $x \in B_c$ implies $y \in B_c$.

Remark : the application of Proposition B.6 is somewhat simplified when the retraction Ret, in addition to being regular and geodesic, is contractive and uniformly geodesic. Here, contractive means that

$$\|\Phi_x(v)\|_x \leq \|v\|_x \qquad \text{for } x \in M \text{ and } v \in T_xM \tag{97}$$

In this case, it is always possible to put $\beta_c = 1$ and $H_c = L_c$, where L_c is the same constant as in Proposition B.4. Uniformly geodesic means there exists $\delta \geq 0$, such that

$$\|\Phi_x(v) - v\|_x \leq \delta \|v\|_x^3 \qquad \text{for } x \in M \text{ and } v \in T_xM \tag{98}$$

In this case, it is always possible to put $\delta_c = \delta$ (independent of c and even of f). ■

The widely-used projection retractions for spheres, unitary groups and Grassmann manifolds, are examples of contractive, uniformly geodesic regular retractions [13] (see Sections 1.5 and 1.6). Here is another example, for positive-definite matrices.

Example : let $M = \mathrm{P}(N)$, the space of symmetric positive-definite $N \times N$ matrices, equipped with its usual affine-invariant metric [6]. For $x \in \mathrm{P}(N)$, the tangent space $T_x\mathrm{P}(N)$ is identified with the space $\mathrm{S}(N)$ of symmetric $N \times N$ matrices. Then, recall the Riemannian exponential map $\mathrm{Exp}_x(v) = x\exp(x^{-1}v)$ (where exp denotes the matrix exponential), and consider the retraction $\mathrm{Ret}_x(v) = x + v + (1/2)vx^{-1}v$. The point of using this retraction is that the eigenvalues of $\mathrm{Ret}_x(v)$ will always be greater than $1/2$. In addition, it is a contractive, uniformly geodesic regular retraction. ■

B.4 Riemannian gradient descent

Let $f : M \to \mathbb{R}$ be a C^2 function, and Ret $: TM \to M$ a retraction. Together, these yield the Riemannian gradient descent scheme, where $\mu \in (0,1]$ is called the step-size,

$$x^{t+1} = \mathrm{Ret}_{x^t}(-\mu \mathrm{grad} f(x^t)) \qquad t = 0, 1, \ldots \tag{99}$$

Assume f has a compact (non-empty) sublevel set B_c, and choose a constant $L_c \geq 0$ as in Proposition B.4. Also, assume Ret is a contractive, uniformly geodesic regular retraction, as in the remark after Proposition B.6. Specifically, Ret verifies (97) and (98), for some constant $\delta \geq 0$. This allows for Ret = Exp, the Riemannian exponential, in which case $\delta = 0$. Now, the following lemma is a direct consequence of (96) in Proposition B.6.

Lemma B.1. *Under the two assumptions just described, on f and* Ret*, if $x^0 \in B_c$, then*

$$\frac{1}{2} - (L_c/2)\mu - (\delta\|\mathrm{grad} f\|_{B_c}^2)\mu^2 \geq 0 \Longrightarrow f(x^{t+1}) \leq f(x^t) - (\mu/2)\|\mathrm{grad} f(x^t)\|_{x^t}^2 \quad (100)$$

for all $t \geq 0$, so that $x^t \in B_c$ for all $t \geq 0$. Here, $\|\mathrm{grad} f\|_{B_c}$ is the maximum of $\|\mathrm{grad} f(x)\|_x$, taken over $x \in B_c$.

This lemma immediately yields the convergence of the Riemannian gradient descent scheme (99), to the stationary points of f in B_c. Here, μ_c^* is the infimum of $\mu \in (0, 1]$ such that $\frac{1}{2} - (L_c/2)\mu - (\delta\|\mathrm{grad} f\|_{B_c}^2)\mu^2 < 0$.

Proposition B.7. *Under the assumptions on f and* Ret*, made in Lemma B.1, if $\mu \leq \mu_c^*$, then $x^0 \in B_c$ implies the sequence (x^t) generated by (99) converges to the set of stationary points of f, in the sublevel set B_c.*

The proof of this proposition is straightforward. From Lemma B.1, if $\mu \leq \mu_c^*$, then $x^0 \in B_c$ implies $x^t \in B_c$ and $(\mu/2)\|\mathrm{grad} f(x^t)\|_{x^t}^2 \leq f(x^t) - f(x^{t+1})$ for all $t \geq 0$. Adding these inequalities, for $t = 0, \ldots, T$,

$$(\mu/2)\sum_{t=0}^{T} \|\mathrm{grad} f(x^t)\|_{x^t}^2 \leq f(x^t) - f(x^{T+1})$$

Then, since f is bounded below on the compact set B_c, the series $\sum_{t=0}^{\infty} \|\mathrm{grad} f(x^t)\|_{x^t}^2$ must converge. Finally, compactness of B_c ensures every subsequence of (x^t) has a further subsequence that converges to a stationary point of f in B_c.

Proposition B.7 holds without any convexity assumptions, made on the function f. Consider now the case of a strictly convex, and then of a strongly convex f.

B.4.1 Strictly convex case

Now, assume the function f is strictly convex on some strongly convex subset A of M. Moreove, assume f has a compact (non-empty) sublevel set $B_c \subset A$. Then, choose a constant $L_c \geq 0$ as in Proposition B.4, and let the retraction Ret be as in Lemma B.1.

Assume that f has a unique minimum at $x^* \in B_c$. Let R be the radius of the smallest ball $B(x^*, R)$ such that $B_c \subset B(x^*, R)$, D the maximum of $\|\mathrm{grad} f(x)\|_x$ for $x \in B(x^*, R)$, and $R_c = R + D$. The following Lemma ensures that (99) "contracts the distance to x^*".

Lemma B.2. *Assume that the sectional curvatures of M lie in the interval $[-\kappa_{\min}^2, \kappa_{\max}^2]$, and that $R_c < \mathrm{inj}(x^*)$. If $x^0 \in B_c$ and $\mu \leq \mu_c^*$ (with μ_c^* as in Proposition B.7), then*

$$1 - \kappa(R_c)L_c\mu - (2\delta R_c\|\mathrm{grad} f\|_{B_c})L_c\mu^2 \geq 0 \Longrightarrow d(x^{t+1}, x^*) \leq d(x^t, x^*) \quad (101)$$

where $\kappa(R_c) = \max\{\kappa_{\min} R_c \coth(\kappa_{\min} R_c), |\kappa_{\max} R_c \cot(\kappa_{\max} R_c)|\}$ and $\|\mathrm{grad} f\|_{B_c}$ is the maximum of $\|\mathrm{grad} f(x)\|_x$ for $x \in B_c$.

Remark : $\mathrm{inj}(x^*)$ denotes the injectivity radius of M at x^* [14]. The condition that $R_c < \mathrm{inj}(x^*)$ guarantees the x^t stay away from the cut locus $\mathrm{Cut}(x^*)$. This condition is introduced because the distance function $x \mapsto d(x, x^*)$ is not differentiable on $\mathrm{Cut}(x^*)$, where its Hessian may even diverge to $-\infty$. ■

Let μ_d^* denote the infimum of μ such that $1 - \kappa(R_c)L_c\mu - (2\delta R_c\|\mathrm{grad} f\|_{B_c})L_c\mu^2 < 0$.

Proposition B.8. *Under the same assumptions of Lemma B.2, if* $\mu \leq \min\{\mu_c^*, \mu_d^*\}$*, then*

$$f(x^{t+1}) - f(x^*) \leq \frac{2d^2(x^0, x^*)}{\mu(t+1)} \tag{102}$$

for all $t \geq 0$*. In particular, the sequence* (x^t) *converges to* x^**.*

Remark : the quality of the convergence in (102) depends above all on the step-size μ. The smaller this is, the slower the convergence. From the definitions of μ_c^* and μ_d^*, it is clear that there are two reasons why μ would be smaller : a larger constant δ, and a larger curvature (in absolute value) $\kappa_{\min}$. In theory, one can alway make $\delta = 0$ by using the retraction Ret = Exp, but this requires the ability to compute the Riemannian exponential Exp with sufficient accuracy. ■

Remark : the rate of convergence stated in (102) is a partial generalisation of the rate found in [39], for gradient descent in a Euclidean space. In the Euclidean setting, $\delta = 0$ and $\kappa_{\min} = \kappa_{\max} = 0$. It then follows from Proposition B.8, that (102) obtains whenever $\mu \leq 1/L_c$. Essentially, this is Corollary 2.1.2 (Page 81) in [39]. However, note the restriction $\mu \in (0, 1]$, which is necessary in a curved Riemannian manifold. ■

Here is an optimisation problem, which falls under the scope of Proposition B.8.

Example : let M be a Hadamard manifold, with sectional curvatures bounded below by $-\kappa_{\min}^2 \leq 0$. Fix a cutoff parameter $q > 0$, and define

$$V_y(x) = q^2 \left[1 + (d(x,y)/q)^2\right]^{\frac{1}{2}} - q^2 \qquad \text{for } x, y \in M \tag{103}$$

In [13], it was proved that $V_y : M \to \mathbb{R}$ is strictly convex, but not strongly convex, and that it is $(1 + q\kappa_{\min})$-smooth on M.

Now, let π be a probability distribution on M and consider the problem of minimising

$$V_\pi(x) = \int_M V_y(x)\, \pi(dy) \tag{104}$$

Note that V_π is strictly convex (but not strongly convex), and $(1 + q\kappa_{\min})$-smooth on M, because the same is true of each function V_y. In fact, V_π has compact sublevel sets whenever the distribution π has finite first-order moments [13]. In this case, $V_\pi(x)$ is guaranteed to achieve its minimum at some $x^* \in M$. This x^* is called the robust Riemannian barycentre of π (the adjective "robust" comes from the field of robust statistics [40]).

When applying Lemma B.2 and Proposition B.8 to the present example (with $f = V_\pi$), note that $\text{inj}(x^*) = \infty$, since M is a Hadamard manifold, and $L_c = (1 + q\kappa_{\min})$ does not depend on c.

B.4.2 Strongly convex case

Here, assume the function f is α-strongly convex on some strongly convex subset $A \subset M$. Let $B_c \subset A$ be a sublevel set of f (where $c > \inf_x f(x)$). Because f is strongly convex, B_c is compact, and it is possible to choose a constant $L_c \geq 0$, as in Proposition B.4. Then, let μ_c^* be given as in Proposition B.7. As usual, f has a unique minimum at $x^* \in B_c$.

Proposition B.9. *Under the assumptions just described, if $\mu \leq \mu_c^*$ and $x^0 \in B_c$, then*

$$f(x^t) - f(x^*) \leq (1 - \mu\alpha)^t (f(x^0) - f(x^*)) \tag{105}$$

for all $t \geq 0$. In particular, the sequence (x^t) converges to x^*.

The proof of this proposition follows by replacing Inequality (87) into Lemma B.1. Indeed, if $\mu \leq \mu_c^*$ and $x^0 \in B_c$, then (100) in Lemma B.1 immediately implies

$$f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - (\mu/2)\|\mathrm{grad} f(x^t)\|_{x^t}^2 \qquad \text{for } t \geq 0$$

Thus, replacing (87) into the right-hand side,

$$f(x^{t+1}) - f(x^*) \leq (1 - \mu\alpha)(f(x^t) - f(x^*))$$

and (105) can be obtained by induction.

Remark : (105) shows that $f(x^t)$ converges to the minimum $f(x^*)$, at an exponential rate. In practice, this can still be quite slow, if $\mu\alpha$ is very small. Indeed, one should attempt to use μ as large as possible, in order to benefit from the exponential rate (105). From the definition of μ_c^*, one cannot have μ any larger than $1/L_c$, and $\mu = 1/L_c$ is only possible if $\delta = 0$, which corresponds to using Ret = Exp. ■

Example : let π be a probability distribution on a complete Riemannian manifold M, and define

$$E_\pi(x) = \frac{1}{2} \int_M d^2(x, y)\, \pi(dy) \qquad \text{for } x \in M \tag{106}$$

If the support of π is contained in a ball $B(z, R)$, where $R < R_c(M)$ given by (84), then E_π is C^2 on $B(z, R)$, and has a unique global minimum $x^* \in M$, such that $x^* \in B(z, R)$ [4] (x^* is the Riemannian barycentre of π). In addition, if $R < R_c(M)/2$, then E_π is α-strongly convex on $B(z, R)$, with α equal to $2\kappa_{\max} R \cot(2\kappa_{\max} R)$ ($= 1$ if $\kappa_{\max} = 0$).

In this case, it is possible to apply Proposition B.9 to the present example (with $f = E_\pi$). If M has positive sectional curvatures, it is always possible to choose $L_c = 1$. On the other hand, if M has negative sectional curvatures $L_c = 1 + 4\kappa_{\min} R$ always works.

C Proofs for Section B

Proof of Proposition B.3 : write inequality (86) under the equivalent form

$$f(y) - f(x) \geq \langle \mathrm{grad} f(x), \mathrm{Exp}_x^{-1}(y) \rangle_x + (\alpha/2) \|\mathrm{Exp}_x^{-1}(y)\|_x^2$$

With $y = x^*$, this becomes

$$f(x) - f(x^*) \leq -\langle \mathrm{grad} f(x), \mathrm{Exp}_x^{-1}(x^*) \rangle_x - (\alpha/2) \|\mathrm{Exp}_x^{-1}(x^*)\|_x^2$$

or, by completing the square on the righ-hand side,

$$f(x) - f(x^*) \leq -\frac{1}{2\alpha} \|\mathrm{grad} f(x) + \mathrm{Exp}_x^{-1}(x^*)\|_x^2 + \frac{1}{2\alpha} \|\mathrm{grad} f(x)\|_x^2$$

Then, (87) follows immediately, by noting the first term on the righ-hand side is negative.

Proof of Proposition B.4 : the proof employs the notation introduced before the proposition. Let $x \in B_c$ and $v = -\mu \mathrm{grad} f(x)$. Then, note that $\|v\|_x \leq \|\mathrm{grad} f(x)\|_x \leq G$. This implies $\gamma(t) = \mathrm{Exp}_x(tv)$ belongs to B' for all $t \in (0, 1)$. From the definition of L_c, it now follows by (89) that

$$f(y) \leq f(x) + \langle \mathrm{grad} f(x), v \rangle_x + (L_c/2) \|v\|_x^2$$

and, by recalling $v = -\mu \mathrm{grad} f(x)$,

$$f(y) \leq f(x) - \mu(1 - (L_c/2)\mu) \|\mathrm{grad} f(x)\|_x^2$$

Then, (90) follows because $\mu \leq 1/L_c$ implies the expression in parentheses is $\geq 1/2$.

Proof of Proposition B.5 : this is given in [13], Section 1.5.

Proof of Proposition B.6 : assume Ret is a regular geodesic retraction, and let Φ be the corresponding map in (93). Since B_c is compact, there exist $\beta_c, \delta_c \geq 0$ such that

$$\sup \{ \|\Phi_x'(u)\|_{\mathrm{op}} ; x \in B_c \text{ and } u \in T_x M, \|u\|_x \leq \|\mathrm{grad} f(x)\|_x \} \leq \beta_c^{1/2} \tag{107}$$

$$\sup \{ \|\Phi_x'''(u)\|_{\mathrm{op}} ; x \in B_c \text{ and } u \in T_x M, \|u\|_x \leq \|\mathrm{grad} f(x)\|_x \} \leq \delta_c \tag{108}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of the linear map $\Phi_x'(u) : T_x M \to T_x M$, or of the tri-linear map $\Phi_x'''(u) : T_x M \times T_x M \times T_x M \to \mathbb{R}$. In terms of these constants β_c and δ_c,

$$\|\Phi_x(-\mu \mathrm{grad} f(x))\|_x^2 \leq (\beta_c \mu^2) \|\mathrm{grad} f(x)\|_x^2 \qquad \text{for } x \in B_c \tag{109}$$

$$\|\Phi_x(-\mu \mathrm{grad} f(x)) + \mu \mathrm{grad} f(x)\|_x \leq (\delta_c \mu^3) \|\mathrm{grad} f(x)\|_x^3 \qquad \text{for } x \in B_c \tag{110}$$

Furthermore, let B_c be contained in a closed geodesic ball $B = \bar{B}(z, R)$. Denote G the maximum of $\|\mathrm{grad} f(x)\|_x$ taken over $x \in B$, and $B' = \bar{B}(z, R + \beta_c^{1/2} G)$. By compactness of B', there exists $H_c \geq 0$ such that f is H_c-smooth on B'.

Now, in order to prove (95), note that $\gamma(t) = \mathrm{Exp}_x(t \Phi_x(-\mu \mathrm{grad} f(x)))$ belongs to B' for all $t \in (0, 1)$. It follows from (94) that (similarly to (89)),

$$f(y) \leq f(x) + \langle \mathrm{grad} f, \Phi_x(-\mu \mathrm{grad} f(x)) \rangle_x + (H_c/2) \|\Phi_x(-\mu \mathrm{grad} f(x))\|_x^2$$

Then, using (109) and (110),

$$f(y) \leq f(x) - \mu \|\mathrm{grad} f(x)\|_x^2 + (\beta_c H_c/2) \mu^2 \|\mathrm{grad} f(x)\|_x^2 + \delta_c \mu^3 \|\mathrm{grad} f(x)\|_x^4$$

which is the same as (95). Finally, (96) is an immediate consequence of (95).

Proof of Lemma B.1 : (100) can be obtained immediately, upon replacing $\beta_c = 1$, $\delta_c = \delta$ and $H_c = L_c$ into (96).

Proof of Proposition B.7 : the proof has already been summarised, right after the proposition.

Proof of Lemma B.2 : let $L(x) = d^2(x, x^*)/2$. If $R_c < \text{inj}(x^*)$, then $L(x)$ is $\kappa(R_c)$-smooth on $B(x^*, R_c)$ [14]. Note that $x^t \in B_c$ for all $t \geq 0$, because $\mu \leq \mu_c^*$ as in Proposition B.7. Then, from the Taylor expansion (94) of L, and since Ret is contractive ((109) holds with $\beta_c = 1$),

$$L(x^{t+1}) \leq L(x^t) + \langle \text{grad}L(x^t), \Phi_{x^t}(-\mu \text{grad} f(x^t)) \rangle_{x^t} + (\kappa(R_c)/2)\mu^2 \|\text{grad} f(x^t)\|_{x_t}^2$$

However, applying (91) to the third term on the right-hand side, this implies

$$L(x^{t+1}) \leq L(x^t) + \langle \text{grad}L(x^t), \Phi_{x^t}(-\mu \text{grad} f(x^t)) \rangle_{x^t} + \kappa(R_c) L_c \mu^2 (f(x^t) - f(x^*)) \quad (111)$$

Now, consider the second term on the right-hand side, since $\text{grad}L(x^t) = -\text{Exp}_{x^t}^{-1}(x^*)$, this second term is equal to

$$\mu \langle \text{Exp}_{x^t}^{-1}(x^*), \text{grad} f(x^t) \rangle_{x^t} - \langle \text{Exp}_{x^t}^{-1}(x^*), \Phi_{x^t}(-\mu \text{grad} f(x^t)) + \mu \text{grad} f(x^t) \rangle_{x^t} \quad (112)$$

Applying (85) and (110) (with $\delta_c = \delta$, since Ret is uniformly geodesic),

$$(112) \ \leq -\mu(f(x^t) - f(x^*)) + (\delta\mu^3) \|\text{Exp}_{x^t}^{-1}(x^*)\|_{x^t} \|\text{grad} f(x^t)\|_{x^t}^3$$

Using (91) once again, along with $\|\text{Exp}_{x^t}^{-1}(x^*)\|_{x^t} \leq R_c$ and $\|\text{grad} f(x^t)\|_{x^t} \leq \|\text{grad} f\|_{B_c}$,

$$(112) \ \leq -\mu(f(x^t) - f(x^*)) + (2\delta R_c \|\text{grad} f\|_{B_c}) L_c \mu^3 (f(x^t) - f(x^*)) \quad (113)$$

Finally, from (111) and (113),

$$L(x^{t+1}) \leq L(x^t) - \mu \left[1 - \kappa(R_c) L_c \mu - (2\delta R_c \|\text{grad} f\|_{B_c}) L_c \mu^2\right] (f(x^t) - f(x^*))$$

Since $f(x^t) \geq f(x^*)$, whenever the expression in square brackets is positive, one has $L(x^{t+1}) \leq L(x^t)$. However, this directly yields (101).

Proof of Proposition B.8 : note from Lemma B.1 that $\mu \leq \mu_c^*$ implies

$$f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - (\mu/2) \|\text{grad} f(x^t)\|_{x^t}^2$$

On the other hand, note that

$$\|\text{grad} f(x^t)\|_{x^t} \geq \frac{f(x^t) - f(x^*)}{d(x^t, x^*)} \geq \frac{f(x^t) - f(x^*)}{d(x^0, x^*)}$$

where the first inequality follows by applying Cauchy-Schwarz to (85), and the second one from Lemma B.2, since $\mu \leq \mu_d^*$. Letting $\varepsilon(t) = f(x^t) - f(x^*)$, it is now clear that

$$\varepsilon(t+1) \leq \varepsilon(t) - (\mu/2) \left(\varepsilon(t)/d(x^0, x^*)\right)^2$$

so that (102) can be proved by a straightforward induction.

Proof of Proposition B.9 : the proof was summarised after the proposition.

References

[1] M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'I.H.P.*, vol. 10, no. 4, pp. 215–210, 1948.

[2] Y. Cabanes, "Multidimensional complex stationary centered gaussian regressive time series classification: Application for audio and dar clutter machine learning in hyperbolic and siegel spaces," Ph.D. dissertation, University of Bordeaux, 2021.

[3] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017.

[4] B. Afsari, "Riemannian L^p center of mass : existence, uniqueness and convexity," *Proceedings of the American Mathematical Society*, vol. 139, no. 2, pp. 655–673, 2010.

[5] S. Said and J. H. Manton, "Riemannian barycentres of Gibbs distributions : new results on concentration and convexity," *Information Geometry*, vol. 4, no. 2, 2021.

[6] X. Pennec, "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 1, pp. 127–154, 2006.

[7] G. Cheng and B. C. Vemuri, "A novel dynamic system in the space of SPD matrices with applications to appearance tracking," *SIAM Journal on Imaging Science*, vol. 6, no. 1, pp. 592–615, 2013.

[8] S., L. Bombrun, Y. Berthoumieu, and J. H. Manton, "Riemannian gaussian distributions on the space of symmetric positive definite matrices," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2153–2170, 2017.

[9] S. Said, H. Hajri, L. Bombrun, and B. C. Vemuri, "Gaussian distributions on Riemannian symmetric spaces : statistical learning with structured covariance matrices," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 752–772, 2018.

[10] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Parameters estimate of Riemannian gaussian distribution in the manifold of covariance matrices," in *Sensor Array and Multichannel Signal Processing*.

[11] L. Santilli and M. Tierz, "Riemannian gaussian distributions, random matrix ensembles and diffusion kernels," *Nuclear Physics B*, vol. 973, 2021.

[12] S. Heuveline, S. Said, and C. Mostajeran, "Gaussian distributions on Riemannian symmetric spaces, random matrices, and planar feynman diagrams," *arXiv:2106.08953*, 2021.

[13] S. Said, "Statistical models and probabilistic methods on riemannian manifolds," *arXiv:2101.10855*, 2021.

[14] P. Petersen, *Riemannian geometry (2nd edition)*, Springer Science, 2006.

[15] S. Helgason, *Differential geometry and symmetric spaces.* New York and London: Academic Press, 1962.

[16] A. W. Knapp, *Lie groups, beyond an introduction (2nd edition).* Birkhauser, 2002.

[17] M. L. Mehta, *Random matrices (3rd edition).* Elsevier Ltd., 2004.

[18] C. L. Siegel, "Symplectic geometry," *American Journal of Mathematics*, vol. 65, no. 1, pp. 1–86, 1943.

[19] A. Terras, *Harmonic analysis on symmetric spaces and applications, Vol. II.* Springer-Verlag, 1988.

[20] K. T. Sturm, "Probability measures on metric spaces of nonpositive curvature," *Contemporary mathematics*, vol. 338, pp. 1–34, 2003.

[21] G. Szegö, *Orthogonal Polynomials (1st edition).* American Mathematical Society, 1939.

[22] A. B. J. Kuijlaars and W. Van Assche, "The asymptotic zero distribution of orthogonal polynomials with varying recurrence coefficients," *Journal of Approximation theory*, vol. 99, pp. 167–197, 1999.

[23] P. Deift, *Orthogonal polynomials and random matrices : a Riemann-Hilber approach*, American Mathematical Sociery, 1998.

[24] M. Mariño, *Chern-Simons theory, matrix models, and topological strings*, Oxford University Press, 2005.

[25] E. T. Whittaker and G. N. Watson, *A course of modern analysis (4th edition)*, Cambridge University Press, 1950.

[26] C. Udriste, *Convex functions and optimization methods on Riemannian manifolds.* Springer Science, 1994.

[27] R. O. Roberts and J. S. Rosenthal, "General state-space Markov chains and MCMC algorithms," *Probability Surveys*, vol. 1, pp. 20–71, 2004.

[28] S. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Cambrdge University Press, 2008.

[29] S. F. Jarner and E. Hansen, "Geometric ergodicity of Metropolis algorithms," *Stochastic Processes and Applications*, vol. 58, pp. 341–361, 1998.

[30] I. Chavel, *Riemannian geometry, a modern introduction.* Cambridge University Press, 2006.

[31] R. Bhattacharya and V. Patrangenaru, "Large sample theory of instrinsic and extrinsic sample means on manifolds I," *The annals of statistics*, vol. 31, no. 1, pp. 1–29, 2003.

[32] D. V. Alekseevskij, E. B. Vinberg, and A. S. Solodovnikov, *Geometry of spaces of constant curvature (EMS vol. 29).* Springer-Verlag, 1993.

[33] V. I. Bogachev, *Measure Theory, Volume I.* Springer-Verlag, 2007.

[34] E. S. Meckes, *The random matrix theory of the classical compact groups.* Cambridge University Press, 2019.

[35] W. S. Kendall, "Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence," *Proceedings of the London Mathematical Society*, vol. 61, no. 2, pp. 371–406, 1990.

[36] J. M. Lee, *Introduction to smooth manifolds (2nd edition)*, Springer Science, 2012.

[37] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition," in *Machine Learning and Knowledge Discovery in Databases*, 2016.

[38] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds.* Princeton University Press, 2008.

[39] Y. Nesterov, *Lectures on convex optimization.* Springer Switzerland, 2018.

[40] P. J. Huber and E. M. Ronchetti, *Robust statistics (2nd edition).* Wiley-Blackwell, 2009.