The Riemannian Barycentre as a Proxy for Global Optimisation

Salem Said[1(✉)] and Jonathan H. Manton[2]

[1] Laboratoire IMS (CNRS 5218), Université de Bordeaux, Bordeaux, France
salem.said@u-bordeaux.fr
[2] Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, Australia
j.manton@ieee.org

Abstract. Let M be a simply-connected compact Riemannian symmetric space, and U a twice-differentiable function on M, with unique global minimum at $x^* \in M$. The idea of the present work is to replace the problem of searching for the global minimum of U, by the problem of finding the Riemannian barycentre of the Gibbs distribution $P_T \propto \exp(-U/T)$. In other words, instead of minimising the function U itself, to minimise $\mathcal{E}_T(x) = \frac{1}{2}\int d^2(x,z)P_T(dz)$, where $d(\cdot,\cdot)$ denotes Riemannian distance. The following original result is proved: if U is invariant by geodesic symmetry about x^*, then for each $\delta < \frac{1}{2}r_{cx}$ (r_{cx} the convexity radius of M), there exists T_δ such that $T \leq T_\delta$ implies $\mathcal{E}_T$ is strongly convex on the geodesic ball $B(x^*, \delta)$, and x^* is the unique global minimum of $\mathcal{E}_T$. Moreover, this T_δ can be computed explicitly. This result gives rise to a general algorithm for black-box optimisation, which is briefly described, and will be further explored in future work.

Keywords: Riemannian barycentre · Black-box optimisation · Symmetric space

It is common knowledge that the Riemannian barycentre $\bar{x}$, of a probability distribution P defined on a Riemannian manifold M, may fail to be unique. However, if P is supported inside a geodesic ball $B(x^*, \delta)$ with radius $\delta < \frac{1}{2}r_{cx}$ (r_{cx} the convexity radius of M), then $\bar{x}$ is unique and also belongs to $B(x^*, \delta)$. In fact, Afsari has shown this to be true, even when $\delta < r_{cx}$ (see [1,2]).

Does this statement continue to hold, if P is not supported inside $B(x^*, \delta)$, but merely concentrated on this ball? The answer to this question is positive, assuming that M is a simply-connected compact Riemannian symmetric space, and $P = P_T \propto \exp(-U/T)$, where the function U has unique global minimum at $x^* \in M$. This is given by Proposition 2, in Sect. 2 below.

Proposition 2 motivates the main idea of the present work: the Riemannian barycentre $\bar{x}_T$ of P_T can be used as a proxy for the global minimum x^* of U. In general, $\bar{x}_T$ only provides an approximation of x^*, but the two are equal if U is invariant by geodesic symmetry about x^*, as stated in Proposition 3, in Sect. 4 below.

F. Nielsen and F. Barbaresco (Eds.): GSI 2019, LNCS 11712, pp. 657–664, 2019.
https://doi.org/10.1007/978-3-030-26980-7_68

The following Sect. 1 introduces Proposition 2, which estimates the Riemannian distance between $\bar{x}_T$ and x^*, as a function of T.

1 Concentration of the Barycentre

Let P be a probability distribution on a complete Riemannian manifold M. A (Riemannian) barycentre of P is any global minimiser $\bar{x} \in M$ of the function

$$\mathcal{E}(x) = \frac{1}{2} \int_M d^2(x,z) P(dz) \quad \text{for } x \in M \tag{1}$$

The following statement is due to Karcher, and was improved upon by Afsari [1,2]: *if P is supported inside a geodesic ball $B(x^*, \delta)$, where $x^* \in M$ and $\delta < \frac{1}{2} r_{cx}$ (r_{cx} the convexity radius of M), then $\mathcal{E}$ is strongly convex on $B(x^*, \delta)$, and P has a unique barycentre $\bar{x} \in B(x^*, \delta)$.*

On the other hand, the present work considers a setting where P is not supported inside $B(x^*, \delta)$, but merely concentrated on this ball. Precisely, assume P is equal to the Gibbs distribution

$$P_T(dz) \;=\; (Z(T))^{-1} \exp\left[-\frac{U(z)}{T}\right] \mathrm{vol}(dz);\; T > 0 \tag{2}$$

where $Z(T)$ is a normalising constant, U is a C^2 function with unique global minimum at x^*, and vol is the Riemannian volume of M. Then, let $\mathcal{E}_T$ denote the function $\mathcal{E}$ in (1), and let $\bar{x}_T$ denote any barycentre of P_T.

In this new setting, it is not clear whether $\mathcal{E}_T$ is differentiable or not. Therefore, statements about convexity of $\mathcal{E}_T$ and uniqueness of $\bar{x}_T$ are postponed to the following Sect. 2. For now, it is possible to state the following Proposition 1. In this proposition, $d(\cdot,\cdot)$ denotes Riemannian distance, and $W(\cdot,\cdot)$ denotes the Kantorovich (L^1-Wasserstein) distance [3,4]. Moreover, $(\mu_{\min}, \mu_{\max})$ is any open interval which contains the spectrum of the Hessian $\nabla^2 U(x^*)$, considered as a linear mapping of the tangent space $T_{x^*}M$.

Proposition 1. *Assume M is an n-dimensional compact Riemannian manifold with non-negative sectional curvature. Denote δ_{x^*} the Dirac distribution at x^*. The following hold,*
(i) *for any $\eta > 0$,*

$$W(P_T, \delta_{x^*}) < \frac{\eta^2}{(4 \operatorname{diam} M)} \implies d(\bar{x}_T, x^*) < \eta \tag{3}$$

(ii) *for $T \leq T_o$ (which can be computed explicitly)*

$$W(P_T, \delta_{x^*}) \leq \sqrt{2}\, (\pi/2)^{n-1}\, B_n^{-1}\, (\mu_{\max}/\mu_{\min})^{n/2}\, (T/\mu_{\min})^{1/2} \tag{4}$$

where $B_n = B(1/2, n/2)$ in terms of the Beta function.

Proposition 1 is motivated by the idea of using $\bar{x}_T$ as an approximation of x^*. Intuitively, this requires choosing T so small that P_T is sufficiently close to δ_{x^*}. Just how small a T may be required is indicated by the inequality in (4). This inequality is optimal and explicit, in the following sense.

It is optimal because the dependence on $T^{1/2}$ in its right-hand side cannot be improved. Indeed, by the multi-dimensional Laplace approximation (see [5], for example), the left-hand side is equivalent to $\mathrm{L} \cdot T^{1/2}$ (in the limit $T \to 0$). While this constant L is not tractable, the constants appearing in Inequality (4) depend explicitly on the manifold M and the function U. In fact, this inequality does not follows from the multi-dimensional Laplace approximation, but rather from volume comparison theorems of Riemannian geometry [6].

In spite of these nice properties, Inequality (4) does not escape the curse of dimensionality. Indeed, for fixed T, its right-hand side increases exponentially with the dimension n (note that B_n decreases like $n^{-1/2}$). On the other hand, although T_o also depends on n, it is typically much less affected by dimensionality, and decreases slower that n^{-1} as n increases.

2 Convexity and Uniqueness

Assume now that M is a simply-connected, compact Riemannian symmetric space. In this case, for any T, the function $\mathcal{E}_T$ turns out to be C^2 throughout M. This results from the following lemma.

Lemma 1. *Let M be a simply-connected compact Riemannian symmetric space. Let $\gamma : I \to M$ be a geodesic defined on a compact interval I. Denote $\mathrm{Cut}(\gamma)$ the union of all cut loci $\mathrm{Cut}(\gamma(t))$ for $t \in I$. Then, the topological dimension of $\mathrm{Cut}(\gamma)$ is strictly less than $n = \dim M$. In particular, $\mathrm{Cut}(\gamma)$ is a set with volume equal to zero.*

Remark: *The assumption that M is simply-connected cannot be removed, as the conclusion does not hold if M is a real projective space.*

The proof of Lemma 1 uses the structure of Riemannian symmetric spaces, as well as some results from topological dimension theory [7] (Chapter VII). The notion of topological dimension arises because it is possible $\mathrm{Cut}(\gamma)$ is not a manifold. The lemma immediately implies, for all t,

$$\mathcal{E}_T(\gamma(t)) \;=\; \frac{1}{2}\int_M d^2(\gamma(t), z) P_T(dz) \;=\; \frac{1}{2}\int_{M-\mathrm{Cut}(\gamma)} d^2(\gamma(t), z) P_T(dz)$$

Then, since the domain of integration avoids the cut loci of all the $\gamma(t)$, it becomes possible to differentiate under the integral. This is used in obtaining the following (the assumptions are the same as in Lemma 1).

Corollary 1. *For $x \in M$, let $G_x(z) = \nabla f_z(x)$ and $H_x(z) = \nabla^2 f_z(x)$, where f_z is the function $x \mapsto \frac{1}{2} d^2(x, z)$. The following integrals converge for any T*

$$G_x \;=\; \int_{M-\mathrm{Cut}(x)} G_x(z)\, P_T(dz); \quad H_x \;=\; \int_{M-\mathrm{Cut}(x)} H_x(z)\, P_T(dz)$$

and both depend continuously on x. Moreover,

$$\nabla \mathcal{E}_T(x) = G_x \quad \text{and} \quad \nabla^2 \mathcal{E}_T(x) = H_x \tag{5}$$

so that $\mathcal{E}_T$ is C^2 throughout M.

With Corollary 1 at hand, it is possible to obtain Proposition 2, which is concerned with the convexity of $\mathcal{E}_T$ and uniqueness of $\bar{x}_T$. In this proposition, the following notation is used

$$f(T) = (4/\pi)\,(\pi/8)^{n/2}\,(\mu_{\max}/T)^{n/2}\exp\left(-U_\delta/T\right) \tag{6}$$

where $U_\delta = \inf\{U(x) - U(x^*)\,;\, x \notin B(x^*, \delta)\}$ for positive δ. The reader may wish to note the fact that $f(T)$ decreases to 0 as T decreases to 0.

Proposition 2. *Let M be a simply-connected compact Riemannian symmetric space. Let κ^2 be the maximum sectional curvature of M, and $r_{cx} = \kappa^{-1}\frac{\pi}{2}$ its convexity radius. If $T \leq T_o$ (see (ii) of Proposition 1), then the following hold for any $\delta < \frac{1}{2} r_{cx}$.*
***(i)** for all x in the geodesic ball $B(x^*, \delta)$,*

$$\nabla^2 \mathcal{E}_T(x) \geq \mathrm{Ct}(2\delta)\,(1 - \mathrm{vol}(M) f(T)) - \pi A_M f(T) \tag{7}$$

where $\mathrm{Ct}(2\delta) = 2\kappa\delta\cot(2\kappa\delta) > 0$ and $A_M > 0$ is a constant given by the structure of the symmetric space M.
***(ii)** there exists T_δ (which can be computed explicitly), such that $T \leq T_\delta$ implies $\mathcal{E}_T$ is strongly convex on $B(x^*, \delta)$, and has a unique global minimum $\bar{x}_T \in B(x^*, \delta)$. In particular, this means $\bar{x}_T$ is the unique barycentre of P_T.*

Note that *(ii)* of Proposition 2 generalises the statement due to Karcher [1], which was recalled in Sect. 1.

3 Finding T_o and T_δ

Propositions 1 and 2 claim that T_o and T_δ can be computed explicitly. This means that, with some knowledge of the Riemannian manifold M and the function U, T_o and T_δ can be found by solving scalar equations. The current section gives the definitions of T_o and T_δ.

In the notation of Proposition 1, let $\rho > 0$ be small enough, so that,

$$\mu_{\min}\, d^2(x, x^*) \;\leq\; 2\,(U(x) - U(x^*)) \;\leq\; \mu_{\max}\, d^2(x, x^*)$$

whenever $d(x, x^*) \leq \rho$, and consider the quantity

$$f(T, m, \rho) \;=\; (2/\pi)^{1/2}\;(\mu_{\max}/T)^{m/2}\;\exp\left(-U_\rho/T\right)$$

where U_ρ is defined as in (6). Note that $f(T, m, \rho)$ decreases to 0 as T decreases to 0, for fixed m and ρ. Now, it is possible to define T_o as

$$T_o \;=\; \min\left\{T_o^1,\, T_o^2\right\} \qquad \text{where} \tag{8}$$

$$T_o^1 = \inf\left\{T > 0 \,:\, f(T, n-2, \rho) \,>\, \rho^{2-n}\, A_{n-1}\,\right\}$$
$$T_o^2 = \inf\left\{T > 0 \,:\, f(T, n+1, \rho) \,>\, (\mu_{\max}/\mu_{\min})^{n/2}\, C_n\right\}$$

Here, $A_n = E|X|^n$ for $X \sim N(0,1)$, and $C_n = \omega_n\, A_n/(\operatorname{diam} M \times \operatorname{vol} M)$, where ω_n is the surface area of a unit sphere S^{n-1}.

With regard to Proposition 2, define T_δ as follows,

$$T_\delta \;=\; \min\left\{T_\delta^1,\, T_\delta^2\right\} - \varepsilon \tag{9}$$

for some arbitrary $\varepsilon > 0$. Here, in the notation of (4), (6) and (7),

$$T_\delta^1 = \inf\left\{T \le T_o \,:\, \sqrt{2\pi}\,(T/\mu_{\min})^{1/2} \,>\, \delta^2\,(\mu_{\min}/\mu_{\max})^{n/2}\, D_n\right\}$$
$$T_o^2 = \inf\left\{T \le T_o \,:\, f(T) \,>\, \mathrm{Ct}(2\delta)\,(\mathrm{Ct}(2\delta)\operatorname{vol} M + \pi A_M)^{-1}\,\right\}$$

where $D_n = (2/\pi)^{n-1}\, B_n/(4 \operatorname{diam} M)$.

4 Black-Box Optimisation

Consider the problem of searching for the unique global minimum x^* of U. In black-box optimisation, it is only possible to evaluate $U(x)$ for given $x \in M$, and the cost of this evaluation precludes numerical approximation of derivatives. Then, the problem is to find x^* using successive evaluations of $U(x)$ (hopefully, as few of these evaluations as possible).

Here, a new algorithm for solving this problem is described. The idea of this algorithm is to find $\bar{x}_T$ using successive evaluations of $U(x)$, in the hope that $\bar{x}_T$ will provide a good approximation of x^*. While the quality of this approximation is controlled by Inequalities (3) and (4) of Proposition 1, in some cases of interest, $\bar{x}_T$ is exactly equal to x^*, for correctly chosen T, as in the following proposition 3.

To state this proposition, let s_{x^*} denote geodesic symmetry about x^* (see [7]). This is the transformation of M, which leaves x^* fixed, and reverses the direction of geodesics passing through x^*.

Proposition 3. *Assume that U is invariant by geodesic symmetry about x^*, in the sense that $U \circ s_{x^*} = U$. If $T \le T_\delta$ (see (ii) of Proposition 2), then $\bar{x}_T = x^*$ is the unique barycentre of P_T.*

Proposition 3 follows rather directly from Proposition 2. Precisely, by *(ii)* of Proposition 2, the condition $T \le T_\delta$ implies $\mathcal{E}_T$ is strongly convex on $B(x^*, \delta)$, and $\bar{x}_T \in B(x^*, \delta)$. Thus, $\bar{x}_T$ is the unique stationary point of $\mathcal{E}_T$ in $B(x^*, \delta)$. But, using the fact that U is invariant by geodesic symmetry about x^*, it is possible to prove that x^* is a stationary point of $\mathcal{E}_T$, and this implies $\bar{x}_T = x^*$. The two following examples verify the conditions of Proposition 3.

Example 1. Assume $M = \mathrm{Gr}(k, \mathbb{C}^n)$ is a complex Grassmann manifold. In particular, M is a simply-connected, compact Riemannian symmetric space. Identify M with the set of Hermitian projectors $x : \mathbb{C}^n \to \mathbb{C}^n$ such that $\mathrm{tr}(x) = k$, where tr denotes the trace. Then, define $U(x) = -\mathrm{tr}(C\,x)$ for $x \in \mathrm{Gr}(k, \mathbb{C}^n)$, where C is a Hermitian positive-definite matrix with distinct eigenvalues. Now, the unique global minimum of U occurs at x^*, the projector onto the principal k-subspace of C. Also, the geodesic symmetry s_{x^*} is given by $s_{x^*} \cdot x = r_{x^*} x\, r_{x^*}$, where $r_{x^*} : \mathbb{C}^n \to \mathbb{C}^n$ denotes reflection through the image space of x^*. It is elementary to verify that U is invariant by this geodesic symmetry.

Example 2. Let M be a simply-connected, compact Riemannian symmetric space, and U_o a function on M with unique global minimum at $o \in M$. Assume moreover that U_o is invariant by geodesic symmetry about o. For each $x^* \in M$, there exists an isometry g of M, such that $x^* = g \cdot o$. Then, $U(x) = U_o(g^{-1} \cdot x)$ has unique global minimum at x^*, and is invariant by geodesic symmetry about x^*.

Example 1 describes the standard problem of finding the principal subspace of the covariance matrix C. In Example 2, the function U_o is a known template, which undergoes an unknown transformation g, leading to the observed pattern U. This is a typical situation in pattern recognition problems.

Of course, from a mathematical point of view, Example 2 is not really an example, since it describes the completely general setting where the conditions of Proposition 3 are verified. In this setting, consider the following algorithm.

Description of the algorithm:

– input : $T \leq T_\delta$ % to find such T, see Section 3
$Q(x, dz) = q(x, z)\mathrm{vol}(dz)$ % symmetric Markov kernel
$\hat{x}_0 = z_0 \in M$ % initial guess for x^*

– iterate : for $n = 1, 2, \ldots$
(1) sample $z_n \sim q(z_{n-1}, z)$
(2) compute $r_n = 1 - \min\{1, \exp[(U(z_{n-1}) - U(z_n))/T]\}$
(3) reject z_n with probability r_n % then, $z_n = z_{n-1}$
(4) $\hat{x}_n = \hat{x}_{n-1} \#_{\frac{1}{n}} z_n$ % see definition (10) below

– until : $\hat{x}_n$ does not change sensibly

– output : $\hat{x}_n$ % approximation of x^*

The above algorithm recursively computes the Riemannian barycentre $\hat{x}_n$ of the samples z_n generated by a symmetric Metropolis-Hastings algorithm (see [8]). Here, The Metropolis-Hastings algorithm is implemented in lines (1)--(3). On the other hand, line (4) takes care of the Riemannian barycentre. Precisely, if $\gamma : [0, 1] \to M$ is a length-minimising geodesic connecting $\hat{x}_{n-1}$ to z_n, let

$$\hat{x}_{n-1} \#_{\frac{1}{n}} z_n \;=\; \gamma\,(1/n) \tag{10}$$

This geodesic γ need not be unique.

The point of using the Metropolis-Hastings algorithm is that the generated z_n eventually sample from the Gibbs distribution P_T. The convergence of the distribution P_n of z_n to P_T takes place exponentially fast. Indeed, it may be inferred from [8] (see Theorem 8, Page 36)

$$\|P_n - P_T\|_{TV} \leq (1 - p_T)^n \tag{11}$$

where $\|\cdot\|_{TV}$ is the total variation norm, and $p_T \in (0,1)$ verifies

$$p_T \leq (\mathrm{vol}(M)) \inf_{x,z} q(x,z) \exp(-\sup_x U(x)/T)$$

so the rate of convergence is degraded when T is small.

Accordingly, the intuitive justification of the above algorithm is the following. Since the z_n eventually sample from the Gibbs distribution P_T, and the desired global minimum x^* of U is equal to the barycentre $\bar{x}_T$ of P_T (by Proposition 3), then the barycentre $\hat{x}_n$ of the z_n is expected to converge to x^*.

It should be emphasised that, in the present state of the literature, there is no rigorous result which confirms this convergence $z_n \to x^*$. It is therefore an open problem, to be confronted in future work.

For a basic computer experiment, consider $M = S^2 \subset \mathbb{R}^3$, and let

$$U(x) = -P_9(x^3) \quad \text{for } x = (x^1, x^2, x^3) \in S^2 \tag{12}$$

where P_9 is the Legendre polynomial of degree 9 [9]. The unique global minimiser of U is $x^* = (0,0,1)$, and the conditions of Proposition 3 are verified, since U is invariant by reflection in the x^3 axis, which is geodesic symmetry about x^*.

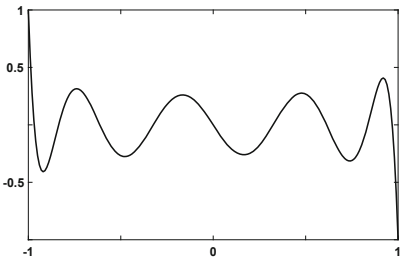

Fig. 1. graph of $-P_9(x^3)$

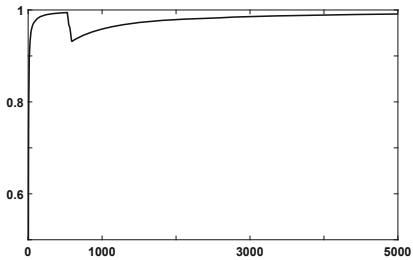

Fig. 2. $\hat{x}_n^3$ versus n

Figure 1 shows the dependence of $U(x)$ on x^3, displaying multiple local minima and maxima. Figure 2 shows the algorithm overcoming these local minima and maxima, and converging to the global minimum $x^* = (0,0,1)$, within $n = 5000$ iterations. The experiment was conducted with $T = 0.2$, and the Markov kernel Q obtained from the von Mises-Fisher distribution (see [10]). The initial guess $\hat{x}_0 = (0,0,-1)$ is not shown in Fig. 2.

In comparison, a standard simulated annealing method offered less robust performance, which varied considerably with the choice of annealing schedule.

Proofs

The proofs of all results stated in this work are detailed in the extended version, available online: https://arxiv.org/abs/1902.03885

References

1. Karcher, H.: Riemannian centre of mass and mollifier smoothing. Commun. Pure. Appl. Math. **30**(5), 509–541 (1977)
2. Afsari, B.: Riemannian L^p center of mass: existence, uniqueness, and convexity. Proc. Am. Math. Soc. **139**(2), 655–673 (2010)
3. Kantorovich, L.V., Akilov, G.P.: Functional Analysis, 2nd edn. Pergamon Press, Oxford (1982)
4. Villani, C.: Optimal Transport, Old and New, 2nd edn. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-71050-9
5. Wong, R.: Asymptotic Approximations of Integrals. Society for Industrial and Applied Mathematics, Philadelphia (2001)
6. Chavel, I.: Riemannian Geometry: A Modern Introduction. Cambridge University Press, Cambridge (2006)
7. Helgason, S.: Differential Geometry, Lie Groups, and Symmetric Spaces. American Mathematical Society, Providence (1978)
8. Roberts, G.O., Rosenthal, J.S.: General state space Markov chains and MCMC algorithms. Probab. Surv. **1**, 20–71 (2004)
9. Beals, R., Wong, R.: Special Functions: A Graduate Text. Cambridge University Press, Cambridge (2010)
10. Mardia, K.V., Jupp, P.E.: Directional Statistics. Academic Press Inc., London (1972)