Shuffling for incremental gradient descent to compute the Riemannian barycenter

Estelle Massart [1] **Julien M. Hendrickx** [1] **P.-A. Absil** [1]

We propose a variant of the incremental gradient algorithm, endowed with a deterministic shuffling strategy, to average a set of data points lying on a Riemannian manifold. The shuffling strategy is shown numerically to improve the convergence rate (with respect to other orderings, including random permutations) for two manifolds of interest, namely, the sphere and the manifold of symmetric positive definite (SPD) matrices.

Averaging a set of data points is a crucial task in many problems, including data denoising, classification and clustering (e.g., with the k-means algorithm). In several applications, the data belong to a curved space (more specifically, a Riemannian manifold), which motivates the development of efficient averaging tools on these spaces. Examples include DTI (diffusion tensor imaging) images denoising (Pennec et al., 2006) and electroencephalogram decoding (Massart & Chevallier, 2017), for which the data are SPD matrices. Another example is clustering for image and video-based recognition. In that case, the data belong to the Grassmann and Stiefel manifolds (Turaga et al., 2011).

A common definition of a mean on a Riemannian manifold $\mathcal{M}$ is the Riemannian barycenter of the data (Karcher, 1977):

$$M = \operatorname*{argmin}_{X \in \mathcal{M}} \frac{1}{2} \sum_{i=1}^{N} f_i(X), \tag{1}$$

where $f_i := \delta^2(X, A_i)$ is the squared Riemannian distance to A_i, and $A_1 \dots, A_N$ are the data points. A classical way to optimize a cost function of the form (1) consists in decreasing each term successively; this approach is known as *incremental gradient descent* (IGD). The terms can be either visited in a deterministic order or in a stochastic order. We endow the IGD algorithm with a deterministic shuffling algorithm, to improve the convergence rate, and illustrate the results on the sphere and the set of SPD matrices.

[1]ICTEAM Institute, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. Supported by EOS Project no 3046816. Correspondence to: Estelle Massart <estelle.massart@uclouvain.be>.

Working Notes of a ICML/FAIM 2018 Workshop, Stockholm, Sweden, 2018.

1. Incremental gradient descent update

Given a sequence of indices $p_0, p_1, p_2, \dots$, with $p_i \in \{1, \dots, N\}$ for all i, and an initial iterate X^0, chosen here to be equal to A_1, the incremental gradient descent update for problem (1) reads:

$$X^{k+1} = \mathrm{Exp}_{X^k}\left(-t_k \operatorname{grad} f_{p_k}(X^k)\right) = X^k \#_{t_k} A_{p_k}, \tag{2}$$

where the notation $A \#_t B$ stands for the endpoint geodesic between A and B, evaluated at parameter t. In the Euclidean case, if the $N-1$ first iterations consist in drawing geodesics (straight lines) towards the points $A_2, \dots, A_N$ (i.e., $p_0, \dots, p_{N-2}$ is a permutation of $\{2, 3, \dots, N\}$), with a steplength chosen as $t_k = 1/(k+2)$, the barycenter (i.e., arithmetic mean) is equal to iterate X^{N-1}, and the IGD algorithm converges after $N-1$ steps. To our knowledge, it is not clear if there exists orderings and steplengths leading to a finite convergence on a general Riemannian manifold.

2. Shuffling algorithm

In (Massart et al., 2017), we propose a shuffling strategy to improve the convergence rate of IGD on the set of SPD matrices, based on the following observation. Numerical results on this manifold indicate that starting the algorithm with the order $p_0 = 2, p_1 = 3, \dots p_{N-2} = N$ tends to overemphasize the last data points considered (i.e., A_N, $A_{N-1}, \dots$) in the resulting iterate X^{N-1}. More specifically, the distance separating the iterate X^{N-1} and the last data points is on average smaller than the one separating the barycenter from those points. We have recently observed the opposite behavior (i.e., the last data points are underemphasized) on the sphere, which suggests that this property may be related to the curvature of the manifold. Indeed, the manifold of SPD matrices, with its classical affine-invariant metric, has non-positive curvature, while the sphere has curvature equal to one. Numerical results indicate that the proposed deterministic shuffling provides on average, both on the sphere and the set of SPD matrices, a faster convergence than other orderings, including random permutations.

References

Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

Massart, E. and Chevallier, S. Inductive means and sequences applied to online classification of EEG. In *Proceedings of the 3rd International Conference on Geometric Science of Information*, pp. 763–770, Paris, 2017.

Massart, E. M., Hendrickx, J. M., and Absil, P.-A. Matrix geometric means based on shuffled inductive sequences. *Linear Algebra and its Applications*, 2017.

Pennec, X., Fillard, P., and Ayache, N. A Riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

Turaga, P., Veeraraghavan, A., Srivastava, A., and Chellappa, R. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, 2011.