Wald Space for Phylogenetic Trees

Jonas Lueg[1](✉), Maryam K. Garba[2], Tom M. W. Nye[3],
and Stephan F. Huckemann[1]

[1] Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences,
Georg-August-Universität, Göttingen, Germany
jonas.Lueg@uni-goettingen.de

[2] Department of Mathematical Sciences, Bayero University, Kano, Nigeria

[3] School of Mathematics, Statistics and Physics, Newcastle University,
Newcastle upon Tyne, UK

Abstract. The recently introduced wald space models phylogenetic trees from an evolutionary perspective. We show that it is a stratified space and propose algorithms to compute geodesics. In application we compute a Fréchet mean of three trees of different topologies that is fully resolved, unlike in BHV-space. Both, preliminary results on geodesics and on means suggest that wald space features less stickiness than BHV-space, making it an alternative model for statistical investigations.

1 Introduction

Phylogenetic trees reflect biological species' evolution. They are built from genetic variation over a set of taxa. Curiously, building them for the same set of taxa, from different genes, however, often result in fundamentally different trees, e.g. Rokas et al. (2003). This generates a call for statistics, for instance averaging over different trees while controlling their uncertainty. Also, this is a call for geometry, designing suitable spaces of trees that are both, biologically meaningful and numerically tractable.

A seminal model has been proposed twenty years ago by Billera et al. (2001), abbreviated as the BHV-model. It has the favorable property of being a Riemann stratified space of globally nonpositive curvature, thus admitting unique geodesics and unique Fréchet means. Additionally, since it is locally flat, an abundance of successful algorithms have been developed for their computation that suffer only from inherent combinatorial complexity, e.g. Owen (2011); Bačák (2014); Miller et al. (2015); Brown and Owen (2018).

While this model is mathematically intriguing, more recently new models have been developed with geometries more closely reflecting stochastic biological fundamentals of gene mutations, e.g. Moulton and Steel (2004); Shiers et al. (2016); Garba et al. (2020). In Garba et al. (2018), metrics for phylogenetic trees based on the information geometry of the two-state and four-state model

Acknowledging DFG HU 1575/7, DFG GK 2088, DFG SFB 1465 and the Niedersachsen Vorab of the Volkswagen Foundation.

© Springer Nature Switzerland AG 2021

F. Nielsen and F. Barbaresco (Eds.): GSI 2021, LNCS 12829, pp. 710–717, 2021.
https://doi.org/10.1007/978-3-030-80209-7_76

were proposed (four states because gene entries are taken from one of the four nucleotide bases). This study was continued in Garba et al. (2020, 2021) and - as a further simplification - a continuous model has been proposed with moments matching those of the two-state model.

In this contribution, we briefly review the definition of our new *wald space* (cf. Garba et al. (2020)) and propose algorithms to compute geodesics and Fréchet means. On the one hand, the wald space is geometrically more challenging. It is a stratified space that is isometrically embedded in the space of positive symmetric $N \times N$ matrices $\mathcal{P}$ (where $N \in \mathbb{N}$ is the number of taxa) equipped with the well known affine invariant geometry of globally nonpositive curvature – hence the need for algorithms as sophisticated as those of the BHV-space. On the other hand, we believe it is biologically more meaningful than the BHV-space. For example, in BHV-space the distance of two different trees with edge lengths becoming arbitrary large diverges to infinity. In wald space, such trees converge to the completely disconnected forest, a member of the wald space, along with other forests. Hence these two trees become more and more similar. Simulations and data analyses reveal advantage of wald space: degenerate trees seem to be less *sticky* (sticky means have degenerate limiting distributions) in wald space than in BHV-space, cf. Hotz et al. (2013); Huckemann et al. (2015); Barden et al. (2013, 2018), thus more easily allowing for statistical inference.

Wald space was first proposed at the Oberwolfach workshop 1804 (2018) in the black forest which is the *Schwarzwald* in German.

2 Wald Space

Let $N \in \mathbb{N}$ denote the number of taxa. A *phylogenetic forest* (F, ℓ) is

(i) a forest $F = (V, E)$ with a finite number of *vertices* V, undirected *edges* E such that any two vertices $u, v \in V$ are connected by at most one edge denoted by $\{u, v\}$ and *labeled* vertices $L = \{1, \dots, N\} \subseteq V$, where $v \in V \setminus L$ implies that $\deg(v) \geq 3$,
(ii) with a mapping $\ell \colon E \to (0, \infty)$.

Two phylogenetic forests are equivalent, $(F_1, \ell_1) \sim (F_2, \ell_2)$, if their label sets agree $L_1 = L = L_2$ and if there is a graph isomorphism $f \colon V_1 \to V_2$ such that

(i) $f(u) = u$ for all $u \in L$, and
(ii) $\ell_1(\{u, v\}) = \ell_2(\{f(u), f(v)\})$ for all $\{u, v\} \in E_1$.

Definition 1. *Every equivalence class $W = [F, \ell]$ is called a* wald *and all equivalence classes form the* wald space *$\mathcal{W}$, its geometric structure is defined further below. Disregarding the edge lengths map ℓ, every equivalence class of forests F with regards to (i) above, is a* wald topology*. For a given wald $W = [F, \ell]$, the* grove *of W is $\mathcal{W}_W$ which comprises all $W' = [F', \ell'] \in \mathcal{W}$ where F' and F have the same wald topology.*

In the following, for any connected $u, v \in V$, $E(u,v)$ is the set of edges along the unique path connecting u and v. For $u = v$, we set any sum over $E(u,u)$ equal zero.

With this notation, the map ϕ sending $W = [F, \ell]$ to the $N \times N$ matrix with coordinate entry at $u, v \in L$,

$$\big(\phi(W)\big)_{uv} = \big(\phi([F,\ell])\big)_{uv} := \begin{cases} \exp\Big(-\sum_{e\in E(u,v)} \ell(e)\Big), & \text{if } u \text{ and } v \text{ are connected,} \\ 0, & \text{else,} \end{cases} \tag{1}$$

is well defined and maps $\mathcal{W}$ injectively into the set of symmetric positive $N \times N$ matrices $\mathcal{P}$, cf. Garba et al. (2020).

Recall from Garba et al. (2020, 2021) that the affine invariant Riemannian metric on $\mathcal{P}$ corresponds to the Fisher information geometry for zero-mean nondegenerate N-dimensional Gaussians induced by tree-indexed Gaussian processes, a continuous generalisation of the two-state model. This metric has the advantage of turning $\mathcal{P}$ into a Riemannian manifold of global nonpositive curvature (e.g. Lang (1999)), guaranteeing unique geodesics and unique Fréchet means (e.g. Sturm (2003)). The squared distance induced on $\mathcal{P}$ is given by

$$d^2_{\mathcal{P}}(P,Q) = \mathrm{Tr}\Big[\log\Big(\sqrt{P}^{-1}Q\sqrt{P}^{-1}\Big)^2\Big] = \sum_{i=1}^{N}\log(\mu_i)^2,$$

where $\sqrt{P}$ is the unique positive definite square root of P and μ_i are the eigenvalues of $P^{-1}Q$.

Definition 2. *The metric $d_{\mathcal{W}}$ of the* wald space *is the pullback of $d_{\mathcal{P}}$ under ϕ, which is given for $W_1, W_2 \in \mathcal{W}$ by*

$$d_{\mathcal{W}}(W_1,W_2) = \inf_{\substack{\gamma\colon [0,1]\to\mathcal{W} \\ \phi\circ\gamma \text{ cont. path,} \\ \gamma(0)=W_1,\gamma(1)=W_2}} L_{d_{\mathcal{P}}}(\phi\circ\gamma),$$

where $L_{d_{\mathcal{P}}}(\gamma)$ is the length of the path γ measured in $d_{\mathcal{P}}$. If no such path exists, we set $d_{\mathcal{W}}(W_1,W_2) = \infty$.

As previously noted, trees with edge lengths ℓ tending to infinity move infinitively far apart in the BHV geometry. In the wald geometry the distance between these trees goes to zero. This is reflected in the following reparametrization $W = [F, \lambda]$ with $\lambda := 1 - \exp(-\ell)$, recasting (1) as

$$\big(\phi(W)\big)_{uv} = \big(\phi([F,\lambda])\big)_{uv} := \begin{cases} \prod_{e\in E(u,v)}\big(1-\lambda(e)\big), & \text{if } u \text{ and } v \text{ are connected,} \\ 0, & \text{else.} \end{cases}$$

In particular, if $W = [F, \lambda]$, $F = (V, E)$, has $|E|$ edges, vectorizing $\lambda \in (0,1)^{|E|}$, we have the following identification for the grove of W:

$$\mathcal{W}_W \cong (0,1)^{|E|}\,.$$

Theorem 1. *1. For every wald $W = [F, \lambda]$, $F = (V, E)$ with grove $\mathcal{W}_W$, the mapping $(0, 1)^{|E|} \cong \mathcal{W}_W \xrightarrow{\phi} \mathcal{P}$ is an embedding.*
2. If $W = [F, \lambda]$ with a fully resolved (i.e. binary) tree F then $\mathcal{W}_W$ is an open subset of $\mathcal{W}$.

Proof. cf. Lueg et al. (2021).

In consequence, $\mathcal{W}$ is a stratified space with strata given by groves. As BHV-space can be viewed as a subset of wald space, cf. Garba et al. (2020), BHV-orthants are subsets of groves. In contrast to BHV-space, groves are not only connected to the star stratum (trees without interior edges), they are also connected to forest strata including the completely disconnected forest (consisting of N isolated vertices, no edges), which lies on the boundary of the star stratum.

3 Geodesics in Wald Space

We propose different algorithms to compute geodesics between two fully resolved trees W_1 and W_2, where Algorithm 4 is only applicable if W_1 and W_2 lie in a common grove $\mathcal{W}_W$. Dropping the embedding map ϕ, we consider wald space $\mathcal{W}$ as a subset of the ambient space $\mathcal{P}$. To this end, for $P, Q \in \mathcal{P}$, denote the unique geodesic between P and Q by $\gamma_{P,Q} \colon [0, 1] \to \mathcal{P}$, the Riemann exponential and logarithm by $\mathrm{Exp}_P^{(\mathcal{P})} \colon T_P\mathcal{P} \to \mathcal{P}$ and $\mathrm{Log}_P^{(\mathcal{P})} \colon \mathcal{P} \to T_P\mathcal{P}$, respectively, the orthogonal tangent space projection by $\pi_W \colon T_P\mathcal{P} \to T_W\mathcal{W}$ and define the projection $\pi \colon \mathcal{P} \to \mathcal{W}, P \mapsto \pi(P) := \mathrm{argmin}_{W \in \mathcal{W}}\, d_{\mathcal{P}}(P, W)$, where π is only well-defined for $P \in \mathcal{P}$ close enough to $\mathcal{W}$. The following is a very simple but naive algorithm.

Algorithm 1 (Naive Projection (NP)). Given $3 \le n \in \mathbb{N}$, $W_1, W_2 \in \mathcal{W}$, for $i = 1, \ldots, n$ compute

(1) $X_i = \pi\big(\gamma_{W_1,W_2}(\frac{i-1}{n-1})\big)$.

Return $(X_1, \ldots, X_n)$.

The next algorithm makes small (approximately geodesic) steps and successively takes the geodesic from the newest point to the destination (note the X_{i-1} and Y_{i-1} in the subscript in the update step).

Algorithm 2 (Successive Projection (SP)). Given $3 \le n \in \mathbb{N}$, $W_1, W_2 \in \mathcal{W}$, set $X_1 := W_1$ and $Y_1 := W_2$. For $i = 2, \ldots, \lfloor \frac{n}{2} \rfloor$, do

(1) $X_i := \pi\big(\gamma_{X_{i-1},Y_{i-1}}(\frac{1}{n-i+1})\big)$ and
(2) $Y_i := \pi\big(\gamma_{Y_{i-1},X_{i-1}}(\frac{1}{n-i+1})\big)$.

If n is even, return $(X_1, \ldots, X_{\lfloor \frac{n}{2} \rfloor}, Y_{\lfloor \frac{n}{2} \rfloor}, \ldots, Y_1)$.
If n is odd, set $Z := \pi\big(\gamma_{X_{\lfloor \frac{n}{2} \rfloor},Y_{\lfloor \frac{n}{2} \rfloor}}(\frac{1}{2})\big)$ and return $(X_1, \ldots, X_{\lfloor \frac{n}{2} \rfloor}, Z, Y_{\lfloor \frac{n}{2} \rfloor}, \ldots, Y_1)$.

The following two algorithms are inspired by Schmidt et al. (2006). They update a given path iteratively and perform a straightening of the path, eventually leading to a geodesic (cf. Figs. 1–4).

Algorithm 3 (Extrinsic Path Straightening (EPS)). Let $3 \leq n \in \mathbb{N}$, $m \in \mathbb{N}$, $W_1, W_2 \in \mathcal{W}$ and suppose $(X_1, \dots, X_n)$ is a path in $\mathcal{W}$ from W_1 to W_2. For $j = 1, \dots, m$, do

(1) for $i = 2, \dots, n-1$ compute $V_i = \frac{1}{2}\big(\mathrm{Log}_{X_i}(X_{i-1}) + \mathrm{Log}_{X_i}(X_{i+1})\big)$ and
(2) update $(X_2, \dots, X_{n-1})$: for $i = 2, \dots, n-1$ compute $X_i \coloneqq \pi\big(\mathrm{Exp}^{(\mathcal{P})}_{X_i}(V_i)\big)$.

Return $(X_1, \dots, X_n)$.

Exploiting the manifold structure of groves, for two walds $W_1, W_2 \in \mathcal{W}_{[F]}$ with the same fully resolved tree F, we change Algorithm 3 slightly and thus avoid using the projection.

Algorithm 4 (Intrinsic Path Straightening (IPS)). Let $3 \leq n \in \mathbb{N}$, $m \in \mathbb{N}$, $W_1, W_2 \in \mathcal{W}_W$ and suppose $(X_1, \dots, X_n)$ is a path in $\mathcal{W}_W$ from $X_1 \coloneqq W_1$ to $X_n \coloneqq W_2$. For $j = 1, \dots, m$, do

(1) for $i = 2, \dots, n-1$ compute $V_i = \frac{1}{2}\pi_{X_i}\Big(\big(\mathrm{Log}_{X_i}(X_{i-1}) + \mathrm{Log}_{X_i}(X_{i+1})\big)\Big)$ and
(2) update $(X_2, \dots, X_{n-1})$: for $i = 2, \dots, n-1$ compute $X_i \coloneqq \mathrm{Exp}^{(\mathcal{W}_W)}_{X_i}(V_i)$.

Return $(X_1, \dots, X_n)$.

We measure the quality of a proposal $(X_1, \dots, X_n)$, $3 \leq n \in \mathbb{N}$ by its length,

$$L(X_1, \dots, X_n) = \sum_{i=1}^{n-1} d_{\mathcal{P}}(X_i, X_{i+1})$$

and its energy,

$$E(X_1, \dots, X_n) = \frac{1}{2}\sum_{i=1}^{n-1} d_{\mathcal{P}}(X_i, X_{i+1})^2.$$

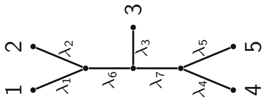

Fig. 1. Tree with edge weights $\lambda^{(1)} = (0.5, \dots, 0.5, 0.1, 0.8)$ and $\lambda^{(2)} = (0.5, \dots, 0.5, 0.9, 0.1)$ for computation of geodesics in Fig. 2.

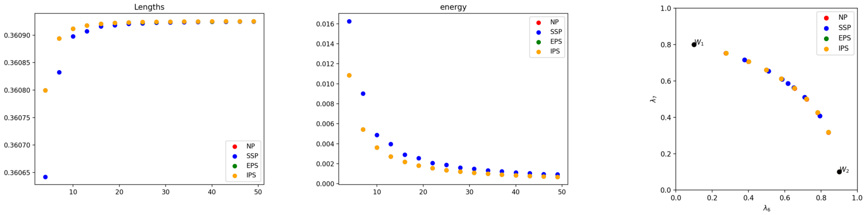

Fig. 2. Length (left) and energy (center) of paths between the two trees from Fig. 1 obtained from the four algorithms for $n = 4, 7, \ldots, 46, 49$. Right: coordinates λ_6, λ_7 of the paths obtained from the four algorithms for $n = 10$. Note that (NP), (IPS), (EPS) almost coincide.

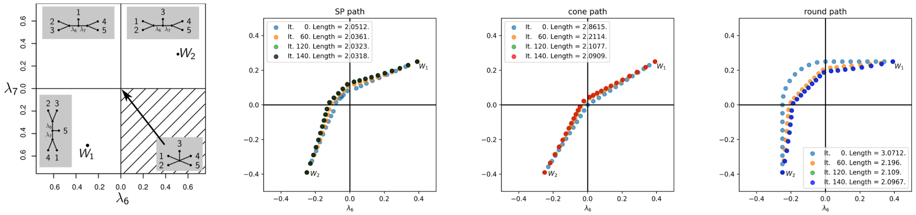

Fig. 3. Left: the coordinate representation (only interior edges) of different neighbouring groves and two walds $W_1, W_2 \in \mathcal{W}$. Second to left to right: Selected iterations of the (EPS) algorithm for different starting paths: the output of the (SP) algorithm, the cone path and a round path, respectively. All paths have $n = 25$ points.

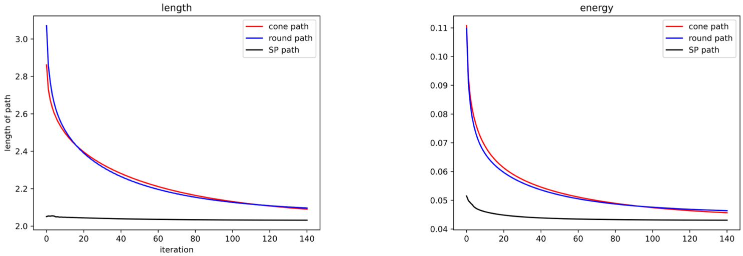

Fig. 4. Left: length of the paths for the iterations of the (EPS) algorithm for different starting paths. Right: energy of the paths for the iterations of the (EPS) algorithm for different starting paths.

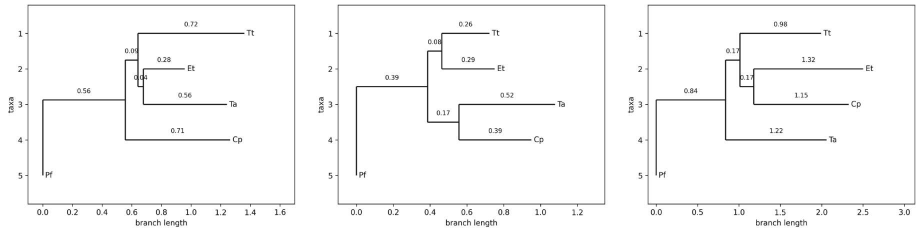

Fig. 5. Three trees $W_1, W_2, W_3 \in \mathcal{W}$ from Nye et al. (2016). Their Fréchet means are depicted in Fig. 6.

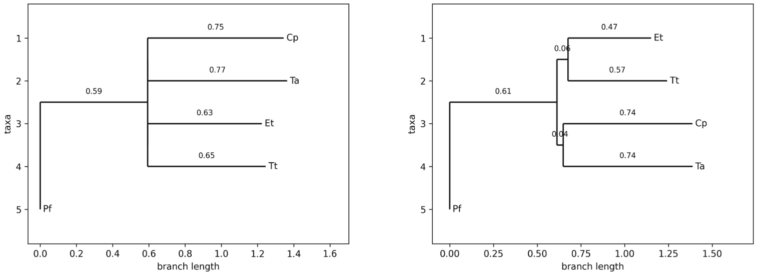

Fig. 6. Fréchet means from the three trees from Fig. 5. Left: in BHV-space, right: in wald space.

4 Comparing Fréchet Means

For illustration, we take $n = 3$ trees $W_1, \ldots, W_n \in \mathcal{W}$ from Nye et al. (2016) depicted in Fig. 5, each of which having $N = 5$ leaves (3 taxa and the root were removed from the original trees for computational tractability). We compute their Fréchet means

$$W^* \in \underset{W \in \mathcal{W}}{\operatorname{argmin}} \sum_{k=1}^{n} d_{\mathcal{W}}\big(W_k, W\big)^2$$

in BHV-space and in wald space, cf. Fig. 6. For computation we use the algorithm of Sturm (2003). In general, the computation of other types of means is also possible (e.g. the Riemannian 1-center, cf. Arnaudon et al. (2013)).

While in BHV-space, the Fréchet mean is unique, in wald space its uniqueness is dubious. For both spaces we have performed 15 iterations after which the final subsequent iterates were less than 0.05 apart, respectively. Remarkably, the mean tree in BHV-space is a star tree. In wald space, however, it is a fully resolved tree.

References

Arnaudon, M., Nielsen, F.: On approximating the Riemannian 1-center. Computational Geometry **46**(1), 93–104 (2013)

Barden, D., Le, H., Owen, M.: Central limit theorems for Fréchet means in the space of phylogenetic trees. Electron. J. Probab **18**(25), 1–25 (2013)

Barden, D., Le, H., Owen, M.: Limiting behaviour of Fréchet means in the space of phylogenetic trees. Annals of the Institute of Statistical Mathematics **70**(1), 99–129 (2016). https://doi.org/10.1007/s10463-016-0582-9

Bačák, M.: Computing Medians and Means in Hadamard Spaces. SIAM Journal on Optimization **24**(3), 1542–1566 (2014)

Billera, L., Holmes, S., Vogtmann, K.: Geometry of the space of phylogenetic trees. Advances in Applied Mathematics **27**(4), 733–767 (2001)

Brown, D. G. and M. Owen (2018, May). Mean and Variance of Phylogenetic Trees. arXiv:1708.00294 [math, q-bio, stat]. arXiv: 1708.00294

Garba, M.K., Nye, T.M., Boys, R.J.: Probabilistic Distances Between Trees. Systematic Biology **67**(2), 320–327 (2018)

Garba, M. K., Nye, T. M. W., Lueg, J., Huckemann, S. F.: Information geometry for phylogenetic trees. Journal of Mathematical Biology **82**(3), 1–39 (2021). https://doi.org/10.1007/s00285-021-01553-x

Garba, M. K., T. M. W. Nye, J. Lueg, and S. F. Huckemann (2021). Information metrics for phylogenetic trees via distributions of discrete and continuous characters. In: Nielsen, F., Barbaresco, F. (Eds.) GSI 2021, LNCS 12829, pp. 701–709 (2021). https://doi.org/10.1007/978-3-030-80209-7_75

Hotz, T., Huckemann, S., Le, H., Marron, J.S., Mattingly, J., Miller, E., Nolen, J., Owen, M., Patrangenaru, V., Skwerer, S.: Sticky central limit theorems on open books. Annals of Applied Probability **23**(6), 2238–2258 (2013)

Huckemann, S., Mattingly, J.C., Miller, E., Nolen, J.: Sticky central limit theorems at isolated hyperbolic planar singularities. Electronic Journal of Probability **20**(78), 1–34 (2015)

Lang, S.: Fundamentals of Differential Geometry. Graduate Texts in Mathematics. Springer-Verlag, New York (1999)

Lueg, J., T. Nye, M. Garba, and S. F. Huckemann (2021). Phylogenetic wald spaces. manuscript

Miller, E., Owen, M., Provan, J.S.: July). Polyhedral computational geometry for averaging metric phylogenetic trees. Advances in Applied Mathematics **68**, 51–91 (2015)

Moulton, V., Steel, M.: Peeling phylogenetic 'oranges'. Advances in Applied Mathematics **33**(4), 710–727 (2004)

Nye, T. M., X. Tang, G. Weyenberg, and Y. Yoshida (2016). Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. arXiv preprint arXiv:1609.03045

Owen, M.: Computing geodesic distances in tree space. SIAM Journal on Discrete Mathematics **25**(4), 1506–1529 (2011)

Rokas, A., B. L. Williams, N. King, and S. B. Carroll (2003, October). Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425(6960), 798–804

Schmidt, Frank R., Clausen, Michael, Cremers, Daniel: Shape Matching by Variational Computation of Geodesics on a Manifold. In: Franke, Katrin, Müller, Klaus-Robert., Nickolay, Bertram, Schäfer, Ralf (eds.) DAGM 2006. LNCS, vol. 4174, pp. 142–151. Springer, Heidelberg (2006). https://doi.org/10.1007/11861898_15

Shiers, N., Zwiernik, P., Aston, J.A., Smith, J.Q.: The correlation space of gaussian latent tree models and model selection without fitting. Biometrika **103**(3), 531–545 (2016)

Sturm, K.: Probability measures on metric spaces of nonpositive curvature. Contemporary mathematics **338**, 357–390 (2003)